Journal of Research in Personality 98 (2022) 104230

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

A direct comparison of the temporal stability and criterion validities of experiential and retrospective global measures of subjective well-being

Nathan W. Hudson [a,*], Richard E. Lucas [b], M. Brent Donnellan [b]

[a] Southern Methodist University, United States
[b] Michigan State University, United States

ARTICLE INFO

Keywords:
Well-being
Emotions
Day reconstruction method

ABSTRACT

Within the past several decades, scholars have expressed concerns regarding the psychometric properties of global, retrospective self-reports of well-being (e.g., life satisfaction scales). This has led to the development of purportedly psychometrically superior experiential measures of well-being, such as the day reconstruction method. However, relatively few studies have directly compared the psychometric properties of global and experiential well-being measures. The present study was a one-month longitudinal design in which we collected up to three measures of (1) global well-being and (2) experiential well-being as measured via the day reconstruction method. These data were used to examine the temporal stability in both types of measures. Moreover, we also examined the criterion-related validity of global and experiential well-being by examining their correlations with theoretically-relevant variables. Results indicated that the majority of variance in all well-being variables was stable across one month—with global life satisfaction being the most stable and experiential negative affect being the least stable. Moreover, in our study, the criterion-related validities for global and experiential well-being were similar. These results seem to affirm the reliability and validity of global measures, and suggest that global and experiential measures of well-being may have similar psychometric properties.

1. Introduction

Subjective well-being is a broad, multifaceted construct that reflects a person's overall evaluation of the quality of their life as a whole. Historically, well-being researchers have focused on two separable subcomponents of this construct (Diener, 1984; Kim-Prieto et al., 2005; Lucas et al., 1996). The first component is an individual's *cognitive* evaluation of the quality of their life, as reflected in an explicit judgment of global life satisfaction. The second component reflects the extent to which an individual typically experiences positive and negative *affect* (e.g., Diener et al., 1985; Watson et al., 1988).

More recently, as can be seen in Fig. 1, researchers have noted that these components can be assessed in at least two different ways. The first is through *global* judgments about one's typical levels of life satisfaction or positive and negative affect. The second assessment technique captures an individual's reports of momentary, in vivo *experiences* of positive and negative emotions (e.g., Kahneman et al., 2004; Shiffman et al., 2008). Importantly, scholars have shown that global and experiential measures of well-being sometimes exhibit distinct patterns of correlations with predictors and outcomes; but it is as yet unclear as to whether these different patterns result from differences in the nature of the constructs being measured or differences in the reliability and validity of the measures used to assess them. Accordingly, the goal of the present report was to compare the psychometric properties of these global and experiential methods of assessing well-being.

1.1. A historical perspective on the assessment of well-being

A defining characteristic of well-being is that it is inherently subjective: People get to decide for themselves whether their lives are evaluated positively or negatively (Diener, 1984). Thus, it is not surprising that well-being has historically been studied using self-report measures (e.g., Cantril, 1967; Diener et al., 1985). This approach, which was initially justified based on face validity and ease of administration of global measures, has also been shown to result in scores with a reasonable degree of reliability and validity (for a review, see Diener et al., 2009).

Despite the initial favorable evidence regarding the psychometric properties of global measures, several researchers have argued that providing an accurate assessment of global well-being may prove

* Corresponding author at: Southern Methodist University, Department of Psychology, P.O. Box 750442, Dallas, TX 72575, United States.
E-mail address: nwhudson@smu.edu (N.W. Hudson).

https://doi.org/10.1016/j.jrp.2022.104230
Received 16 June 2021; Received in revised form 19 June 2021; Accepted 3 April 2022
Available online 6 April 2022

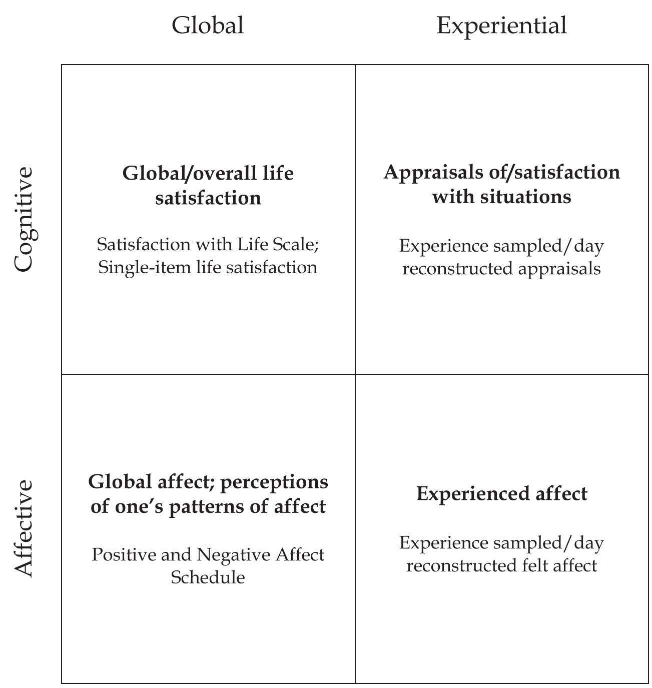

Fig. 1. *Subtypes of Well-Being, Note.* Well-being can be subdivided by two factors (global vs. experiential; cognitive vs. affective) to produce four separate subtypes of well-being. Descriptions of each subcomponent of well-being, along with sample measures thereof, are listed in each cell.

cognitively demanding for respondents (Robinson & Clore, 2002a, 2007). Specifically, these scholars take the position that the process of forming an accurate judgment regarding one's well-being requires mentally aggregating across a considerable amount of relevant information, and can therefore force respondents to rely on potentially inaccurate heuristics and reduce the validity of global measures, at least in some circumstances (Schwarz et al., 1987; Schwarz & Clore, 1983; Schwarz & Strack, 1999). Seemingly supporting this perspective, studies suggest that participants' evaluations of well-being can be partially influenced by theoretically inconsequential factors, such as current mood or fleeting and trivial emotional experiences (e.g., Schwarz & Clore, 1983; Schwarz & Strack, 1999). For example, people's self-reports of overall happiness and satisfaction with their *life as a whole* have been shown to vary as a function of the weather (Schwarz & Clore, 1983) and whether their favorite soccer team had recently won a game (Schwarz et al., 1987). Such findings have prompted conclusions that "reports about happiness and satisfaction with one's life do not necessarily reflect stable inner states" (Schwarz et al., 1987, p. 70), and that "there is 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).[1]

An alternative to global self-reports is to assess an individual's momentary experiences of well-being throughout daily life and then aggregate these reports. The basic idea is that a person high in well-being will report frequent experiences of positive emotions and relatively infrequent experiences of negative emotions. Moreover, reporting one's in vivo emotions should not require intensive cognitive operations, and thus individuals should be able to answer questions about their current (or recent) felt emotions relatively accurately (Robinson & Clore, 2002a, 2002b). Thus, by aggregating emotional experiences, cognitive biases can be minimized and researchers can attain an ostensibly "objective" assessment of a participants' well-being (Kahneman, 1999). This perspective has motivated the development of a number of experiential assessment approaches, including the experience sampling method (ESM; Shiffman et al., 2008) and, more recently, the day reconstruction method (DRM; Kahneman et al., 2004). The DRM is a relatively new approach designed to overcome some of the practical limitations of ESM, such as the relatively high burden ESM places on participants.

Specifically, in the DRM, participants complete a survey in which they reconstruct their entire previous day, listing all activities in which they engaged and rating their affective experiences during those activities. This does not require researchers to purchase costly software or equipment or ask participants to install potentially intrusive phone applications. Moreover, it is less burdensome to participants because it does not require them to carry an electronic paging device or suffer interruptions during their daily routines. Preliminary evidence suggests that, despite the DRM entailing a small degree of retrospective reporting, once aggregated across a day, ESM and DRM measures of experiential well-being correlate strongly with one another (Bylsma et al., 2011; Dockray et al., 2010; Lucas et al., 2016).

An emerging body of research examining the properties of these newer experiential measures (e.g., DRM) has found that global and experiential well-being only moderately correlate with one another (Hudson et al., 2017, 2020). Perhaps more surprisingly, global and experiential measures sometimes show different patterns of correlations with theoretically relevant predictors (Diener & Tay, 2014; Hudson et al., 2016; Kahneman & Deaton, 2010). For example, greater income is associated with higher life satisfaction (to a point) but not daily experiences of happiness (Hudson et al., 2016; Kahneman & Deaton, 2010). This has raised questions about the reasons for these discrepancies.

On the one hand, it may simply be the case that global and experiential well-being reflect fundamentally distinct ways of evaluating one's life—and thus tap different, albeit related constructs (Hudson et al., 2017; Redelmeier & Kahneman, 1996; Robinson & Clore, 2002a). For instance, people's top-down judgments of how frequently they experience positive emotions may not match aggregated measures of their actual, lived moods and feelings because someone could predominantly experience negative affect, yet nevertheless decide that their life as whole is satisfying (or vice versa). Similarly, when evaluating the overall quality of their lives, individuals may idiosyncratically mentally "weight" the importance of different experiences in ways that are difficult for researchers to anticipate. For example, a person might predominantly experience negative affect during most of their waking hours (e.g., at work), but nevertheless conclude that they have high levels of global positive affect because they feel mostly positive emotions "when it matters" (e.g., in social contexts; outside of work). On the other hand, the different patterns of associations between global and experiential well-being may result from differences in psychometric properties of the measures used to assess these different components of well-being (i.e., some measures may be more reliable or valid than others, resulting in stronger associations with predictors and outcomes).

Resolving the debate regarding the assessment of well-being is important for future research because it directly bears on the kinds of decisions researchers will make when designing studies and testing hypotheses. The goal of the current report was to compare global self-report and experiential measures in terms of their convergence, short-term stability coefficients over the course of one month, and criterion-related validity coefficients. The results of our study provide needed data about these two approaches and potentially shed light on the nature of the constructs being assessed by these respective techniques.

1.2. Stability of well-being

There are often debates in the social and behavioral science as to whether specific constructs represent stable, trait-like individual

[1] It important to note that such studies have historically been underpowered to detect typical effect sizes in social psychology, which raises the possibility of inflated effect sizes in this literature (Button et al., 2013). Thus, the case against global self-reports may not be as strong as portrayed (see Yap et al., 2017).

differences or a more transitory states that are heavily influenced by situational factors and fleeting moods (Anusic & Schimmack, 2016; Fraley & Roberts, 2005; Fraley, Vicary, Brumbaugh, & Roisman, 2011; Hudson et al., 2016). Because well-being is supposed to tap one's evaluation of life as a whole, and because relevant life conditions do not typically fluctuate from moment to moment, measures of well-being should be relatively stable over time. Thus, tests of stability can be used to evaluate the validity of any well-being assessment technique. Accordingly, by directly comparing the stability of global self-reports and experiential measures of well-being, it is possible to evaluate how well the two approaches measure stable individual differences.

To that end, it is important to identify benchmarks for interpreting stability coefficients. Existing meta-analyses suggests that test–retest stabilities in global self-reported well-being are approximately $r = 0.70$, 0.60, 0.50, and 0.35 over 1-, 2-, 5-, and 10-year intervals, respectively—asymptotically approaching a lower-bound of approximately 0.20 to 0.35 over increasing long intervals (Anusic & Schimmack, 2016; Schimmack & Oishi, 2005). Thus, given our study's short (one month) duration, we might expect to observe stabilities in global well-being greater than $r = 0.70$. Far fewer studies have examined stability in experiential affect as measured via DRM. Those studies have found that the stabilities in experiential well-being are approximately $r = 0.65$, 0.50, 0.40, and 0.35 over 2-week, 4-week, 1-year, and 2-year test–retest intervals (Hudson et al., 2017, 2020; Krueger & Schkade, 2008). Thus, if anything, we might expect to find *lower* test–retest stability for experiential well-being than global well-being in our study (perhaps approximately $r = 0.50$), potentially suggesting that experiential well-being is *more* influenced by situational factors (e.g., daily events) than are global measures. Notably, lower stability coefficients would suggest that a measure is a potentially less-valid indicator of stable individual differences.

As an important methodological note, people's moment-by-moment experienced emotions are influenced by transitory situational factors and exhibit very low stability (Epstein, 1979)—but when aggregated across time (e.g., across a day), random state-level factors tend to cancel out, and individuals' affective experiences increase in temporal stability, capturing more stable individual differences in trait affect (Diener & Larsen, 1984; Hudson et al., 2017). Although the ideal measure of experiential well-being might be to aggregate individuals' affect across extended periods of time (e.g., weeks or months)—such intensive data collection may not be feasible, especially in large-scale survey work designed to assess well-being amidst other research goals (e.g., Wagner et al., 2007). Moreover, similar logic could be applied to global measures: To the extent they are influenced by transient situational/mood factors (e.g., Schwarz & Clore, 1983; Schwarz & Strack, 1999), repeatedly administering global measures should allow these random situational effects to cancel out, ultimately tapping stable individual differences. Thus, both to inform pragmatic research applications of DRM and to provide a fair comparison of the relative merits of global versus experiential measures of well-being, we collected and aggregated only one day's worth of global and experiential well-being at each time point in our study.

1.3. Criterion-related validity

The stability of scores across time is a critical consideration when evaluating measures of well-being. Another consideration, however, is the extent to which observed scores correlate with theoretically relevant variables. This is often referred to as the measure's *criterion validity* (Cronbach & Meehl, 1955). Intuitively, well-being should correlate with a wide variety of variables including health (Okun et al., 1984), relationship status (Lucas & Dyrenforth, 2006), parental status (Nelson et al., 2013), employment status (Helliwell, 2003), socioeconomic status (Pinquart & Sorensen, 2000), and personality traits such as extraversion and neuroticism (Dyrenforth et al., 2010; Heller et al., 2004; Schimmack et al., 2002; Steel et al., 2008). As it stands, global self-report measures of well-being have been shown to correlate in expected ways with each of these variables, albeit the correlations are often small-to-moderate in size.

In contrast to results for global measures, Diener and Tay (2014) concluded that, "…Feelings assessed by the DRM do not seem to correlate to any degree with important variables such as employment or income…. It is an open question whether the DRM shows significant correlations with important measures that are not collected in the DRM itself" (pp. 259–260). However, the existing database of studies including DRM measures of well-being is sparse and additional research is needed to evaluate their claim—especially studies that simultaneously compares global self-reports and DRM-based approaches. Accordingly, in the present study, we directly compared the extent to which DRM measures of experiential well-being and global measures correlated with personality, health, relationship status, parental status, employment status, and income. To the extent that experiential measures are more valid indicators of well-being than global ones (Kahneman, 1999; Schwarz & Clore, 1983), we should expect the experiential measures to have *greater* criterion-related validities (i.e., greater correlations with these external variables) than do the global measures.

Importantly, it is possible that experiential measures provide a more valid assessment of individuals' underlying well-being—but do so alongside more random noise (or vice versa). In other words, it is possible, for example, that measures of experiential well-being provide a purer (i.e., less systematically biased) assessment of the "true" well-being construct than do global measures, but that they also contain so much measurement error that it is difficult to detect their correlates (Diener & Tay, 2014). To address this possibility, we collected repeated measures of global and experiential well-being and separated variance in the measures across time into the stable, trait-like constructs underlying each measure, in addition to the unique, state variance (which includes measurement error) at each occasion. By separately correlating the trait- and state-components of each type of well-being with external criteria, we were able to test the extent to which the *core constructs* underlying experiential and global well-being predicted external criteria, sans any noise due to random measurement error.

1.4. Overview of the present study

The present study was a one-month longitudinal study in which participants provided up to three waves of data. At each wave, participants completed three global well-being measures—life satisfaction, global positive affect, and global negative affect. Additionally, participants reported their experiential positive and negative affect from the previous day using the DRM. These data were used to evaluate the one-month stabilities in self-reported and experiential well-being. Moreover, we examined the criterion validities of global self-reports and experiential measures by computing participants' latent trait scores (across all waves) for experiential and global well-being and examining the extent to which participants' latent trait well-being correlated with various theoretically relevant criteria.

Finally, as an important ancillary goal, we also used our data to examine to the extent to which two methodological choices might influence the psychometric properties of the DRM. First, when aggregating emotional data across episodes in the DRM, researchers are divided with respect to whether it is better to use raw, unweighted averages, as opposed to averages in which affect ratings are weighted by episode duration (Diener & Tay, 2014; Kahneman et al., 2004). Thus, in the present study, we tested the psychometric properties of both unweighted and duration-weighted experiential affect composites.

Second, we examined whether randomly sampling DRM episodes might produce different results from administering a comprehensive DRM assessment. Namely, in the full DRM, participants list all activities in which they had engaged during the prior day, and rate their affective experiences in *every* episode (Kahneman et al., 2004). This process, however, can take upwards of an hour to complete—which is untenable for many research contexts, as well as large-scale national surveys (e.g., the German Socioeconomic Panel or American Time Use Study). To reduce the time required to complete the DRM, researchers have often adopted the strategy of asking participants to list all activities in which they engaged the prior day, but to rate their affective experiences in only a randomly selected subset of the listed activities (e.g., three randomly selected episodes; Hudson et al., 2020). The psychometric consequences of doing so, however, are not well understood.

To address this issue, in the present study, participants completed the full DRM (i.e., they rated their affective experiences across all reported episodes). Our primary analyses examined composites of participants' affective experiences averaged across all episodes. However, to emulate research in which DRM episodes were randomly sampled, we also created composites of *randomly sampled DRM affect* by randomly selecting three episodes per participant per wave and using only the affective ratings from the sampled episodes to form composites. Thus, we were able to directly compare the psychometric properties of experiential well-being scores computed using the full DRM versus scores based on a randomly sampled subset of DRM episodes.

2. Method

2.1. Participants

The sample was recruited from a list of Michigan residents who had previously participated in at least one wave of the Michigan University State of the State Survey (SOSS; Michigan State University Institute for Public Policy and Social Research, 2015), and who had indicated that they would be interested in receiving invitations to participate in other studies. Specifically, the SOSS is a quarterly, statewide telephone survey of approximately 1000 adult Michigan residents per wave, recruited via stratified random sampling procedures (Pierce, 2016). SOSS participants can opt-in to receive invitations to participate in additional, external studies. The SOSS administration team sent participants who had expressed interest in participating in future research an email invitation to participate in our study, alongside a link to the study website. Participants were offered $20 USD per wave for completing up to three waves, plus a $15 USD bonus for completing all three waves (thus, maximum compensation for completing all waves was $75 USD); participants could opt to receive either Amazon.com credit or a check. All study materials were presented online.

A total of 410 participants responded to the email invitation and provided at least one wave of data.[2] This sample size afforded more than 99% power to detect average-sized effects in personality psychology (equivalent to $r \sim 0.21$; Richard et al., 2003) The final sample at Time 1 was 60% female, with ages ranging from 19 to 92 ($M = 52.77$, $SD = 14.81$). The racial composition of the sample was 86% White, 6% Black, 2% Asian, 2% Native American, and 1% Hispanic. Seventy-four percent of participants indicated they were currently involved in a romantic relationship, 82% had children, and 53% were employed.

At Time 1, participants provided their contact information, and were later contacted and encouraged to provide two additional waves of data, with Time 2 and Time 3 measures collected an average of 17.60 ($SD = 4.84$) and 33.82 ($SD = 6.51$) days after Time 1, respectively. On average, participants provided 2.31 waves of data ($SD = 0.91$), with 326 participants (80%) completing at least two waves. Attrition analyses revealed that only extraversion was related to total waves of data provided ($r = -0.10$, 95% confidence interval [CI] [-0.19, -0.01]). No other study variables, as measured at Time 1, were significantly related to waves of data provided, all $|r|\text{s} \leq 0.06$.

[2] Data from these participants are currently be used for several in-preparation manuscripts examining associations among well-being, time spent with others, activities engaged, personality, and relationship quality.

2.2. Well-being measures

All well-being measures were collected at all three time points.

2.2.1. Experiential well-being

Participants' experiential well-being was measured using a variant of the DRM (Kahneman et al., 2004). Participants were first asked to reconstruct their entire prior day in terms of "scenes" or "episodes" that had occurred. Specifically, participants were given relatively open-ended instructions to divide their prior day's morning, afternoon, and evening into episodes, "name" each episode, and recorded its start and end time. After reconstructing their entire prior day, participants were presented with each episode they had defined, and were asked to (1) select all activities they had performed during the episode from a predetermined list [e.g., commuting, shopping, housework], (2) select with whom they were interacting during the episode from a predetermined list [e.g., spouse, friend, coworker], and (3) rate the extent to which they felt various emotions *during the episode*: happiness, satisfaction, anger, sadness, frustration, and worry. All emotions were rated on a scale from 0 (*not at all*) to 6 (*very much*). We formed daily composites for each emotion by averaging ratings across all episodes with equal weighting (e.g., we formed a *daily happiness* composite by averaging ratings of happiness across all episodes). Because research indicates that positive and negative affect are separable and sometimes independent (Watson et al., 1988), we formed separate composites for *daily positive affect* (an average of daily happiness and satisfaction; Time-1 $\alpha = 0.89$) and *daily negative affect* (an average of daily anger, sadness, frustration, and worry; Time-1 $\alpha = 0.90$).

Duration-Weighted Affect. Our primary analyses used unweighted daily affect composites. That is, participants' affective ratings for each episode within a wave were averaged together with equal weighting to form daily composites. Although this practice has been advocated by scholars due to its parsimony and potential for accurately capturing relevant psychological processes (e.g., happiness in greater numbers of self-defined "episodes" may be more psychologically meaningful than happiness in fewer, albeit longer episodes [e.g., work]; Diener & Tay, 2014), others have argued that well-being should be a literal summation of momentary affect—and thus affective ratings in DRM episodes should be weighted by episode duration when forming overall affective composites (Kahneman et al., 2004). Therefore, in the present study, we also created *weighted daily affect* composites and directly compared their psychometric properties to unweighted composites.[3]

Randomly Sampled Affect. On average, participants reported 12.01 DRM episodes ($SD = 4.96$) per measurement occasion.[4] Although in our study participants rated their affective experiences across all DRM episodes, several prior studies have used abbreviated versions of the DRM in which participants listed all episodes that occurred the prior day, but rated their affective experiences in only three randomly selected episodes (e.g., Hudson et al., 2017). The empirical consequences of randomly sampling episodes in the DRM versus asking participants to report on all episodes are not well understood. Thus, in the present study we also attempted to address this issue by randomly

[3] The weighted composites were created via the following procedure. For each individual emotion (e.g., happiness), participants' rating in each episode was multiplied by the episode duration divided by the total duration of all DRM episodes reported that wave (i.e., percent of total reported time). These products were then summed within waves to produce, for example, a weighted happiness score. The weighted positive and negative emotions were then averaged together to form weighted positive and negative affect composites, respectively.

[4] Participants reported a similar number of episodes across individual waves. A multilevel model revealed that number of reported episodes did not significantly vary as a function of wave number—with participants non-significantly trending toward reporting 0.24 (95% CI [-0.01, 0.49]) more episodes with each passing wave.

selecting three episodes per wave per participant.[5] We then created *sampled negative affect* and *sampled positive affect* composites at each wave by averaging participants' ratings across only the randomly selected episodes. When randomly sampling episodes, duration-weighted composites are less appropriate (see Anusic et al., 2017); so, we therefore weighted all three episodes equally in the sampled affect composites. We repeated this sampling procedure 10 times to create 10 sampled positive and negative affect variables per participant per wave. For all reported results pertaining to sampled affect, we report the average point estimates and confidence interval across the 10 random samples.[6]

2.2.2. Global affect

To measure global affective well-being, participants were asked to rate the extent to which they had felt various emotions over the past two weeks: happy, satisfied, angry, sad, frustrated, and worried. Each emotion was rated from 0 (*almost never*) to 6 (*almost always*). As with experiential well-being, we formed separate composites for *global positive affect* (an average of global happiness and satisfaction; Time-1 $\alpha = 0.84$) and *global negative affect* (an average of global anger, sadness, frustration, and worry; Time-1 $\alpha = 0.79$).

2.2.3. Global life satisfaction

We measured participants' life satisfaction in two ways. First, participants rated their life-satisfaction using the five-item satisfaction with life scale (SWLS; Diener et al., 1985). Items (e.g., "I am satisfied with my life") were rated on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*) and averaged to form a composite (Time-1 $\alpha = 0.90$). Second, we used a single-item life satisfaction measure of the sort that is frequently used in large-scale survey work (e.g., the GSOEP; Wagner et al., 2007). Thus, participants also rated a single-item scale that read, "All things considered, how satisfied are you with your life?" This item was rated from 0 (*completely dissatisfied*) to 10 (*completely satisfied*).

2.3. Criterion variables at Time 1

All criterion variables were measured at only Time 1.

2.3.1. Personality traits

Participants' personality traits were measured using the 20-item Mini-IPIP (Donnellan et al., 2006). The Mini-IPIP contains subscales to measure extraversion (e.g., "I am the life of the party"), agreeableness (e.g., "I sympathize with others' feelings"), conscientiousness (e.g., "I get chores done right away"), emotional stability (the opposite of neuroticism; e.g., "I am relaxed most of the time"), and openness to experience (e.g., "I have a vivid imagination"). All items were rated on a scale from 1 (*very inaccurate*) to 5 (*very accurate*) and averaged together to form composites for each big five personality dimension (αs ranged 0.65 [agreeableness] to 0.77 [extraversion]).

2.3.2. Health

Participants were asked to rate their health using a single-item scale with options ranging from 1 (*poor*) to 5 (*excellent*). Research suggests that, although multi-item health measures tend to correlate more strongly with criterion variables, single item measures of health are nevertheless reliable, valid, and sensitive to within-person variation in health across time (Macias et al., 2015; Wu et al., 2013).

2.3.3. Income

Participants indicated their total annual household income using a scale that ranged from 1 (*less than $10,000*) to 8 (*more than $75,000*). Intermediate values (e.g., 4) represented ranges (e.g., *$20,000 to less than $25,000*). In all analyses, individuals' incomes were coded as the mean of the range (e.g., 4 [$20,000–25,000] was recoded as $22,500). Responses of 1 and 8 were coded as $5,000 and $80,000, respectively.

2.3.4. Other demographics

Participants reported their [1] age, [2] relationship status (1 = in a committed relationship; 0 = not in a committed relationship), [3] parental status (1 = parent; 0 = non-parent), and [4] employment status (1 = employed; 0 = unemployed).

3. Results

Table 1 provides the descriptive statistics and correlations among all study variables at Time 1. Replicating prior research (e.g., Hudson et al., 2017), experiential measures of positive and negative affect were moderately correlated with global self-report measures of positive and negative affect (average $r = 0.54$). We subsequently computed the cross-time latent correlations between all well-being variables. This was accomplished by creating a single latent factor for each well-being variable which captured the shared variance in each measure across all three waves (e.g., a "daily positive affect" latent variable was created with daily positive affect at Time 1, Time 2, and Time 3 as its indicators). The descriptive statistics and correlations for these latent variables are presented in Table 2. On a latent/trait level, experiential and global affect were highly correlated with one another (average $r = 0.77$). This suggests that, in our study, the experiential and global self-report affect measures shared considerable variance.

3.1. Stability in experiential and global self-report well-being measures

We next examined the portion of variance in each well-being measure that was attributable to a constant, trait-like latent variable over the course of the study's duration. To this end, we present two separate, albeit highly related statistical analyses. First, Table 3 contains the manifest test–retest correlations across each time point. Second, we constructed state-trait structural equation models, which capture the percent variance in each measure that was attributable to constant, trait-like dynamics across the study (Kenny & Zautra, 1995). Given the statistical similarity of these approaches, we focus our narrative only on the latter.[7]

As depicted in Fig. 2, in state-trait models, a latent variable is used to capture variance shared across all time points. This latent variable is interpretable as the portion of variance attributable to a constant and unchanging factor over the course of the study (Anusic et al., 2012). The residual terms are therefore interpretable as the influence of transitory state-level factors. Critically, the state-level residuals capture both contextual variance (e.g., "true" state variation in well-being) as well as measurement error. It is possible to use second-order latent trait models (e.g., models in which latent variables are formed from the scale items at

[5] In total, there were 26 data points for which 3 or fewer episodes were provided. In these cases, all episodes provided by the participant at that wave were included in all random samples.

[6] The parameter estimates were generally quite consistent across the 10 random samples. For example, estimates of the trait loading for sampled positive affect (see Table 4) ranged from 0.76 to 0.80 ($SD = 0.01$). Similarly, estimates of the trait loading for sampled negative affect ranged from 0.74 to 0.77 ($SD = 0.02$).

[7] In a pure state-trait model, the model-implied test–retest correlation between any two measurement occasions, i and j, is the product of their trait loadings (i.e., $\lambda_i \times \lambda_j$). When the trait loadings are constrained to be equal across time (as in our study), the implied test–retest correlation between any two time points is therefore simply the trait λ^2. Not coincidentally, the percent variance explained by constant, trait-like dynamics is also λ^2. Thus, the percent trait variance in a state-trait model is conceptually equivalent to the test–retest correlation, constrained to be equal across all possible intervals in the study (e.g., T1-T2, T2-T3, T1-T3).

Table 1
Descriptive Statistics and Correlations at Time 1.

Variable	*M*	*SD*	Correlations															
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Well-being																		
1. Daily positive affect	3.56	1.30	–															
2. Daily negative affect	0.79	0.87	**-0.35**	–														
3. Global positive affect	4.22	0.98	**0.59**	**-0.37**	–													
4. Global negative affect	2.51	1.00	**-0.41**	**0.48**	**-0.58**	–												
5. Multi-item life satisfaction	4.75	1.37	**0.47**	**-0.37**	**0.70**	**-0.54**	–											
6. Single-item life satisfaction	7.33	1.71	**0.49**	**-0.40**	**0.71**	**-0.57**	**0.80**	–										
Correlates																		
7. Age	52.77	14.81	**0.20**	**-0.12**	**0.13**	**-0.24**	**0.13**	**0.18**	–									
8. Extraversion	3.09	0.88	**0.18**	**-0.11**	**0.24**	**-0.11**	**0.18**	**0.17**	-0.02	–								
9. Agreeableness	4.08	0.65	**0.13**	–03	**0.20**	**-0.11**	**0.12**	**0.15**	0.00	**0.16**	–							
10. Conscientiousness	3.74	0.77	**0.22**	**-0.11**	**0.19**	**-0.12**	**0.13**	**0.14**	**0.11**	0.03	**0.15**	–						
11. Emotional stability	3.41	0.83	**0.34**	**-0.39**	**0.48**	**-0.62**	**0.46**	**0.48**	**0.21**	0.03	**0.15**	**0.18**	–					
12. Openness	3.72	0.78	-0.04	-0.03	0.01	0.01	0.01	0.00	-0.03	**0.18**	0.11	-0.08	0.03	–				
13. Health	3.49	0.99	**0.21**	**-0.29**	**0.39**	**-0.25**	**0.46**	**0.43**	-0.01	**0.11**	0.06	**0.21**	**0.31**	0.04	–			
14. In relationship (1 = Yes)	0.74	0.44	0.06	-0.04	**0.11**	-0.02	**0.25**	**0.18**	0.02	0.04	-0.03	0.03	0.00	-0.05	**0.12**	–		
15. Is parent (1 = Yes)	0.82	0.38	**0.16**	-0.01	**0.13**	**-0.10**	**0.18**	**0.15**	**0.30**	0.08	-0.05	**0.13**	0.09	**-0.12**	0.08	**0.25**	–	
16. Is employed (1 = Yes)	0.53	0.50	**-0.13**	-0.04	0.01	0.09	-0.02	-0.05	**-0.48**	-0.06	-0.06	-0.05	-0.08	0.00	0.09	0.05	-0.09	–
17. Income, in thousands	59.15	23.27	0.09	**-0.12**	**0.25**	**-0.14**	**0.35**	**0.25**	**0.10**	0.03	-0.08	0.00	0.06	-0.05	**0.22**	**0.44**	**0.18**	**0.18**

Note. Ninety-five percent confidence intervals for correlations in **boldface** do not contain zero.

each time point) to partition the variance into (1) trait variance, (2) state variance, and (3) measurement error. However, such models require at least two scale items at each time point to converge—and at least three items are recommended (Geiser et al., 2015). Unfortunately, half of our scales contained two or fewer items—and one scale contained only a single item. Thus, to ensure that our parameter estimates could be meaningfully compared across all included measures, we opted for simpler models that use manifest composites at each time point instead of latent factors. Because the same model—with the same constraints—was used across all variables, the estimates from each model can be meaningfully compared to one another. Nonetheless, as a result, our study likely overestimates the amount of state variance in each variable (as measurement error is included in the estimates).[8]

In our models, we standardized the trait and state latent factors (and all reported parameters are standardized). For parsimony—and to allow the trait to be interpreted as the portion of variance that is constant across time—we constrained the trait and state loadings to be equal across time. One particularly useful feature of this model specification is that the squared standardized loadings represent the portion of variance in the measures that is attributable to constant, trait-like factors, as opposed to malleable state-like factors (plus measurement error). Notably, because most of the scales under investigation had two or fewer items, all of the SEMs used manifest composites (rather than latent factors) at each time point.

Table 4 contains the estimates of the variance in each well-being variable that was attributable to state- and trait- factors. Over the one-month study duration, all variables under investigation were remarkably stable—with more than half of the variance in each variable attributable to trait-like latent variables. Life satisfaction—irrespective of whether measured via the SWLS or a single item—exhibited the greatest stability, with more than 80% of its variance attributable to a constant, trait-like latent variable (SWLS $\lambda^2 = 0.86$, 95% CI [0.74, 0.98]; single-item $\lambda^2 = 0.81$, 95% CI [0.70, 0.93]). In contrast, experiential negative affect was the least stable variable, but nonetheless with the majority—57%—of its variance due to trait-like dynamics (95% CI [0.47, 0.69]). Notably, experiential negative affect was also the least stable variable over a period of two years in Hudson and colleagues' (2017) two-year study of a German sample. The remaining variables in our study had similar stabilities—with approximately three quarters of their variance attributable to constant, trait-like latent factors. Thus, our study suggests that both experiential and global well-being tap primarily stable, trait-like constructs. Moreover, with the exception of experiential negative affect, experiential and global measures exhibited similar levels of stability across one month.

3.2. *Criterion validity of experiential and global well-being measures*

Next, we evaluated how the trait- and state-level components of well-

[8] Our state-trait models presented in the main text used manifest composites at each time point (see Fig. 2). Reviewers requested that, when possible, we run models that aggregated items at each time point using latent variables. The benefit of such models is that they separate measurement error from "true" state variance (whereas manifest composites conflate measurement error and state variance). Unfortunately, three of our scales had fewer than three items, and thus we could not form latent variables (with unconstrained, free loadings) at each time point. Thus, in the main text, we present models with manifest composites for all variables (to ensure that results are comparable across variables). However, using latent variables at each time point did not statistically significantly affect estimates of the percent trait variance for global negative affect ($\lambda^2 = 0.80$), experiential negative affect ($\lambda^2 = 0.64$), or the SWLS ($\lambda^2 = 0.89$). With additional constraints (e.g., requiring all items to have identical loadings on the latent factor), it is possible to also construct latent variables with only two items. Using these types of models did not statistically significantly affect estimates of the percent trait variance for global positive affect ($\lambda^2 = 0.86$) or experiential positive affect ($\lambda^2 = 0.76$).

Table 2
Cross-Time Latent Descriptive Statistics and Correlations for Well-Being Variables.

Latent Variable	*M*	*SD*	Latent Correlations								
			1	2	3	4	5	6	7	8	9
1. Daily positive affect	3.55	1.07	–								
2. Daily negative affect	0.79	0.61	**-0.51**	–							
3. Global positive affect	4.22	0.87	**0.82**	**-0.62**	–						
4. Global negative affect	2.51	0.87	**-0.55**	**0.72**	**-0.75**	–					
5. Multi-item life satisfaction	4.75	1.23	**0.67**	**-0.58**	**0.87**	**-0.68**	–				
6. Single-item life satisfaction	7.33	1.51	**0.72**	**0.63**	**0.91**	**-0.74**	**0.93**	–			
7. Weighted positive affect	3.44	1.05	**1.00**	**-0.51**	**0.82**	**-0.54**	**0.65**	**0.70**	–		
8. Weighted negative affect	0.73	0.64	**-0.51**	**1.00**	**-0.61**	**0.77**	**-0.56**	**-0.61**	**-0.54**	–	
9. Sampled positive affect*	3.44	1.14	**1.00**	**-0.47**	**0.80**	**-0.54**	**0.66**	**0.70**	**1.00**	**-0.51**	–
10. Sampled negative affect*	0.60	0.64	**-0.48**	**1.00**	**-0.63**	**0.73**	**-0.57**	**-0.62**	**-0.51**	**1.00**	**-0.47**

Note. Ninety-five percent confidence intervals for correlations in **boldface** do not contain zero.
* Statistics for these variables represent the average estimate across all 10 random samples.

Table 3
Test-Retest Correlations for Well-Being Variables.

Variable	Time 1 – Time 2			–	Time 2 – Time 3			–	Time 1 – Time 3		
		95% CI		–		95% CI		–		95% CI	
	r	LB	UB		*r*	LB	UB		*r*	LB	UB
Daily positive affect	**0.70**	0.64	0.75		**0.77**	0.72	0.81		**0.66**	0.59	0.72
Daily negative affect	**0.54**	0.45	0.61		**0.70**	0.63	0.75		**0.50**	0.40	0.58
Global positive affect	**0.78**	0.74	0.82		**0.81**	0.76	0.84		**0.72**	0.66	0.77
Global negative affect	**0.71**	0.66	0.76		**0.74**	0.68	0.78		**0.69**	0.63	0.75
Multi-item life satisfaction	**0.84**	0.81	0.87		**0.88**	0.85	0.90		**0.87**	0.84	0.90
Single-item life satisfaction	**0.79**	0.74	0.83		**0.84**	0.80	0.87		**0.81**	0.76	0.84
Weighted positive affect	**0.60**	0.52	0.67		**0.73**	0.67	0.78		**0.58**	0.50	0.65
Weighted negative affect	**0.57**	0.49	0.64		**0.70**	0.63	0.75		**0.51**	0.41	0.59
Sampled positive affect*	**0.59**	0.51	0.66		**0.67**	0.60	0.73		**0.56**	0.42	0.60
Sampled negative affect*	**0.54**	0.45	0.61		**0.67**	0.60	0.73		**0.52**	0.42	0.60

Note: CI = confidence interval; LB = lower-bound; UB = upper-bound.
95% CIs for coefficients in **boldface** do not include zero.
* Statistics for these variables represent the average estimate across all 10 random samples.

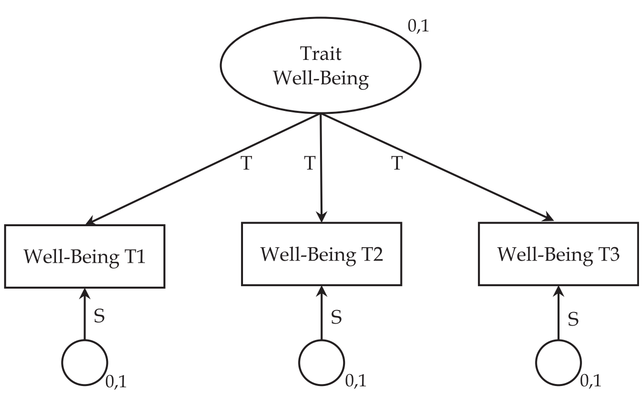

Fig. 2. State-Trait Model of Well-Being.

being were associated with various theoretically relevant correlates. As depicted in Fig. 3, this involved regressing each criterion variable simultaneously onto the stable, trait-like latent component of each well-being variable, as well as the Time 1 state component (because the criteria were only measured at Time 1).[9] Thus, these analyses directly compare the extent to which the *latent-construct* underlying each well-being measure was related to criteria, as opposed to any measurement-specific state variance (e.g., any occasion-specific factors that caused, for example, people to rate their well-being *and* extraversion more highly would be captured as the correlation between health and state well-being, and *not* the correlation between health and trait well-being). We conducted separate analyses for each criterion variable. All reported parameters are standardized.

The results of these analyses are presented in Table 5. It is first worth noting that, with a few exceptions, the criteria were generally only related to latent trait levels of both experiential and global well-being. This suggests that personality traits, for example, were associated with people's typical levels of well-being, rather than concurrent idiosyncratic fluctuations in well-being at a single point in time. In other words, each criterion-variable was more strongly related to stable individual differences in well-being, rather than to random fluctuations that occurred in the same measurement occasion (e.g., "common assessment-occasion variance").

When specifically considering the trait associations in the upper half of Table 5, we found little evidence for differences in the criterion validities of global and experiential well-being. For example, as expected, extraversion was associated with both greater global positive affect (β = 0.27, 95% [0.16, 0.37]) and experiential positive affect (β = 0.20, 95% CI [0.09, 0.31]), and emotional stability was correlated with both global negative affect (β = -0.77, 95% CI [-0.89, −0.65]) and experiential negative affect (β = -0.48, 95% CI [-0.61, −0.35]). Similarly, as expected, people with greater health reported greater positive affect both globally (β = 0.42, 95% CI [0.31, 0.53]) and experientially (β = 0.32, 95% CI [0.21, 0.44]) as well as less negative affect, both globally (β = -0.33, 95% CI [-0.44, −0.21]) and experientially (β = -0.39, 95% CI [-0.52, −0.27]). Generally, the point estimates for the associations with

[9] All models fit well, all CFIs > 0.97, RMSEAs $\leq$ 0.07.

Table 4
Trait- and State-Level Variance in Well-Being.

Well-Being Variable	Trait						State					
		95% CI			95% CI			95% CI			95% CI	
	λ	LB	UB	λ^2	LB	UB	λ	LB	UB	λ^2	LB	UB
Daily positive affect	**0.84**	0.77	0.91	**0.71**	0.60	0.83	**0.54**	0.51	0.57	**0.29**	0.26	0.33
Daily negative affect	**0.76**	0.69	0.83	**0.57**	0.47	0.69	**0.65**	0.62	0.69	**0.43**	0.38	0.47
Global positive affect	**0.88**	0.81	0.95	**0.77**	0.66	0.89	**0.48**	0.45	0.50	**0.23**	0.20	0.25
Global negative affect	**0.84**	0.78	0.91	**0.71**	0.60	0.83	**0.54**	0.51	0.57	**0.29**	0.26	0.32
Multi-item life satisfaction	**0.93**	0.86	0.99	**0.86**	0.74	0.98	**0.38**	0.36	0.40	**0.14**	0.13	0.16
Single-item life satisfaction	**0.90**	0.84	0.97	**0.81**	0.70	0.93	**0.44**	0.41	0.46	**0.19**	0.17	0.21
Weighted positive affect	**0.80**	0.73	0.87	**0.64**	0.53	0.75	**0.60**	0.57	0.64	**0.36**	0.32	0.41
Weighted negative affect	**0.76**	0.69	0.84	**0.58**	0.48	0.70	**0.64**	0.61	0.68	**0.42**	0.37	0.46
Sampled positive affect*	**0.78**	0.71	0.85	**0.60**	0.50	0.71	**0.63**	0.60	0.67	**0.40**	0.36	0.45
Sampled negative affect*	**0.75**	0.68	0.82	**0.57**	0.47	0.68	**0.66**	0.62	0.69	**0.43**	0.38	0.48

Note. CI = confidence interval; LB = lower-bound; UB = upper-bound.
All parameter estimates are standardized; because of how the model is specified, λ^2 represents the proportion of variance in each variable that is attributable to trait- or state-level dynamics; 95% CIs for coefficients in **boldface** do not include zero.
* Statistics for these variables represent the average estimate across all 10 random samples.

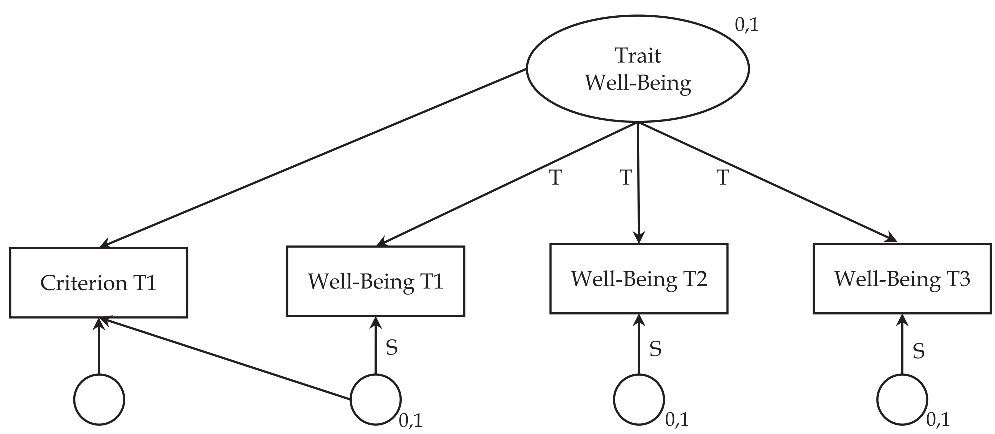

Fig. 3. State and Trait Components of Well-Being Predicting Criterion Variables.

experiential well-being fell within the confidence intervals for the associations with global well-being (and vice versa)—suggesting that global and experiential well-being did not exhibit substantially different patterns of criterion validity. Indeed, the average absolute association between global well-being and the criteria examined ($\beta = 0.20$) was nearly identical to the average absolute association between the criteria and experiential well-being ($\beta = 0.19$).[10]

3.3. Duration-weighted experiential affect

Thus far, our analyses of experiential well-being used composites in which participants' affective ratings within a wave were averaged across all episodes with equal weighting. For our next series of analyses, we examined whether using duration-weighted composites might produce different results. As can be seen in Table 2, on a latent level, the duration-weighted composites were correlated perfectly with the unweighted composites. This may suggest that the only difference between these two weighting methods is random/measurement error. However, as can be seen by comparing the "daily affect" and "weighted affect" rows in Tables 3-4, the duration-weighted positive affect composite was slightly less stable (trait $\lambda^2 = 0.64$, 95% CI [0.53, 0.75]) than was the unweighted positive affect composite (trait $\lambda^2 = 0.71$, 95% CI [0.60, 0.83]). In contrast, the weighted and unweighted negative affect composites were nearly identical in terms of stability (both trait λ^2s = 0.57-0.58).

Finally, using duration-weighted versus unweighted affective composites did not substantially affect the pattern of criterion validities. The average absolute correlation between the criterion variables and experiential affect was identical—$|r| = 0.19$—irrespective of whether the duration-weighted or unweighted DRM composites were used. Thus, neither weighted nor unweighted affective composites had systematically higher criterion correlations. With the exception of duration-weighted and unweighted negative affect predicting openness (respective correlations: $r = 0.06$, 95% CI [-0.07, 0.18]; $r = -0.07$, 95% CI [-0.19, 0.06]; $\Delta r = 0.13$), all individual criterion correlations for weighted versus unweighted affective were within $\Delta r = \pm 0.06$ of each other.

3.4. Randomly sampled experiential affect

Finally, we examined the extent to which forming composites of experiential affect based on three randomly sampled DRM episodes (as opposed to all DRM episodes) influenced our findings. To clarify

[10] Using second-order latent models (for variables for which it was possible) did not significantly change these results. As a representative example, for trait multi-item life satisfaction, the standardized associations with age, extraversion, agreeableness, conscientiousness, emotional stability, openness, health, relationship status, parental status, employment status, and income were 0.10, 0.25, 0.10, 0.16, 0.43, 0.00, 0.45, 0.22, 0.15, -0.01, and 0.32, respectively (compare with the numbers in the top half of Table 5).

Table 5
Regressions Predicting Criteria from Trait and State Components of Well-Being Simultaneously.

	Trait																	
	Daily PA			Daily NA			Global PA			Global NA			Multi-item LS			Single-item LS		
		95% CI			95% CI			95% CI			95% CI			95% CI			95% CI	
Criterion	β	LB	UB	β	LB	UB	β	LB	UB	β	LB	UB	β	LB	UB	β	LB	UB
Age	**0.27**	0.15	0.38	-0.10	-0.23	0.02	**0.16**	0.05	0.27	**-0.33**	-0.45	-0.22	0.10	-0.01	0.20	**0.17**	0.07	0.28
E	**0.20**	0.09	0.31	**-0.20**	-0.32	-0.07	**0.27**	0.16	0.37	**-0.12**	-0.23	-0.01	**0.23**	0.13	0.34	**0.26**	0.15	0.36
A	0.08	-0.03	0.20	0.02	-0.11	0.14	**0.16**	0.05	0.26	-0.03	-0.14	0.08	**0.11**	0.01	0.21	**0.13**	0.03	0.24
C	**0.32**	0.20	0.44	**-0.22**	-0.34	-0.09	**0.18**	0.08	0.29	**-0.13**	-0.24	-0.03	**0.15**	0.05	0.25	**0.21**	0.11	0.31
S	**0.41**	0.29	0.52	**-0.48**	-0.61	-0.35	**0.56**	0.44	0.67	**-0.77**	-0.89	-0.65	**0.49**	0.38	0.59	**0.51**	0.40	0.62
O	-0.03	-0.14	0.08	0.06	-0.07	0.18	0.01	-0.09	0.12	-0.01	-0.12	0.10	0.00	-0.09	0.10	0.00	-0.10	0.11
Health	**0.32**	0.21	0.44	**-0.39**	-0.52	-0.27	**0.42**	0.31	0.53	**-0.33**	-0.44	-0.21	**0.50**	0.40	0.61	**0.46**	0.35	0.57
In Relationship	0.11	-0.01	0.22	-0.10	-0.22	0.03	0.08	-0.03	0.18	0.00	-0.11	0.11	**0.24**	0.13	0.34	**0.15**	0.05	0.26
Is Parent	**0.18**	0.06	0.29	-0.12	-0.25	0.00	**0.14**	0.03	0.24	-0.07	-0.18	0.04	**0.16**	0.06	0.26	**0.12**	0.02	0.22
Is Employed	**-0.15**	-0.26	-0.03	-0.05	-0.18	0.07	-0.03	-0.13	0.08	**0.13**	0.02	0.24	-0.02	-0.12	0.08	-0.06	-0.16	0.04
Income	**0.12**	0.00	0.23	**-0.20**	-0.33	-0.08	**0.26**	0.15	0.37	**-0.15**	-0.26	-0.03	**0.33**	0.22	0.43	**0.23**	0.12	0.34

	State																	
	Daily PA			Daily NA			Global PA			Global NA			Multi-item LS			Single-item LS		
		95% CI			95% CI			95% CI			95% CI			95% CI			95% CI	
Criterion	β	LB	UB	β	LB	UB	β	LB	UB	β	LB	UB	β	LB	UB	β	LB	UB
Age	-0.03	-0.16	0.11	-0.06	-0.20	0.07	-0.01	-0.14	0.13	0.06	-0.07	0.19	0.09	-0.05	0.22	0.06	-0.08	0.19
E	0.03	-0.11	0.16	0.04	-0.09	0.17	0.03	-0.10	0.16	-0.03	-0.16	0.10	-0.07	-0.20	0.07	-0.10	-0.24	0.03
A	0.11	-0.03	0.24	-0.05	-0.19	0.08	**0.14**	0.00	0.27	**-0.18**	-0.32	-0.05	0.05	-0.08	0.19	0.06	-0.07	0.20
C	-0.07	-0.20	0.06	0.07	-0.07	0.20	0.08	-0.05	0.21	-0.02	-0.15	0.11	-0.02	-0.15	0.12	-0.08	-0.22	0.05
S	0.05	-0.09	0.19	-0.05	-0.18	0.09	**0.16**	0.03	0.29	**-0.31**	-0.44	-0.18	**0.14**	0.00	0.27	0.13	0.00	0.26
O	-0.03	-0.16	0.11	-0.09	-0.22	0.05	-0.01	-0.14	0.12	0.04	-0.09	0.17	0.02	-0.12	0.15	-0.01	-0.14	0.13
Health	-0.08	-0.22	0.05	-0.01	-0.14	0.12	0.13	0.00	0.26	0.01	-0.12	0.14	0.11	-0.02	0.25	0.13	0.00	0.26
In Relationship	-0.04	-0.17	0.10	0.04	-0.10	0.17	0.10	-0.03	0.23	-0.04	-0.17	0.09	0.10	-0.03	0.23	0.09	-0.03	0.22
Is Parent	0.03	-0.10	0.17	0.10	-0.03	0.24	0.03	-0.11	0.16	-0.08	-0.22	0.05	0.08	-0.06	0.21	0.10	-0.03	0.24
Is Employed	-0.02	-0.15	0.12	0.01	-0.13	0.14	0.08	-0.06	0.21	-0.04	-0.17	0.09	0.00	-0.13	0.14	0.00	-0.13	0.13
Income	0.00	-0.14	0.13	0.04	-0.10	0.17	0.09	-0.05	0.22	-0.03	-0.16	0.10	**0.15**	0.01	0.28	0.12	-0.02	0.25

Note. PA = positive affect; NA = negative affect; LS = life satisfaction; CI = confidence interval; LB = lower bound; UB = upper bound; E = extraversion; A = agreeableness; C = conscientiousness; S = emotional stability; O = openness. The "In Relationship," "Is Parent," and "Is Employed" variables were dummy coded (0 = no, 1 = yes). All parameter estimates are standardized. 95% CIs for parameters in **boldface** do not include zero.

terminology, we use *sampled affect* to refer to composites of affect from three randomly sampled DRM episodes, and *daily affect* to refer to composites of affect across all DRM episodes. To generate point estimates for sampled affect, we repeated the sampling procedure a total of ten times (and thus created, e.g., 10 sampled positive affect variables). In all tables, we report the *average* parameter estimates and confidence interval bounds across all ten samples.

As can be seen in Table 2, on a latent level, sampled affect correlated perfectly with daily affect ($rs = 1.00$). This is precisely what would be expected given a large enough sample size and true random sampling of episodes. Specifically, from a classical test theory perspective, affect in each episode should reflect a combination of an individual's "true score," as well as truly random error (including both occasion-specific error and measurement error) (Lord & Novick, 1968). Thus, the *theoretical* expected value of affect in any given single episode is simply the person's true, latent affective well-being. Moreover, this is true irrespective of the number of episodes aggregated together; aggregating across increasingly large numbers of episodes does not change the expected value of the composite. Of course, this is true *in practice* only with sufficient sample sizes for random errors to mutually cancel across observations. Our analyses suggest that a sample size of 410 individuals with 3 episode ratings each was sufficient for the latent correlations between latent affect as measured via all versus randomly sampled DRM episodes to reach unity.

Nevertheless, as can be seen in Table 4, sampled positive affect was less stable across time (trait $\lambda^2 = 0.60$, 95% CI [0.50, 0.71]) than was daily positive affect (trait $\lambda^2 = 0.71$, 95% CI [0.60, 0.83]). This, however, was not true for negative affect (both trait λ^2s $= 0.57$, 95% CI [0.47, 0.69]). These findings suggest that sampling fewer DRM episodes per measurement occasion has the potential to less reliably tap stable individual differences in at least experiential positive affect. This likely reflects that, with greater number of DRM episodes rated, random measurement error in the composite tends to mutually cancel to a greater degree, producing better estimates of participants' latent trait experiential affect.

Finally, the correlations between latent trait sampled affect and external criteria were largely identical to the correlations between latent trait daily affect and external criteria. For example, the average absolute correlation between sampled positive affect and the criteria ($\beta = 0.19$) was nearly identical to the average absolute correlation between daily positive affect and the criteria ($\beta = 0.20$), and the maximum difference in any specific criterion-correlation was $\Delta r = 0.03$ (e.g., daily positive affect correlated $\beta = 0.32$ with conscientiousness, whereas sampled positive affect correlated on average $\beta = 0.29$ with conscientiousness).

4. Discussion

Scholars have called for direct comparisons of the psychometric properties of experiential well-being and global well-being (Diener & Tay, 2014). The purpose of the present study was to address this lacuna and directly compare the one-month temporal stability and criterion validity of global self-reported and experiential well-being measured via DRM. In general, our findings converged on the conclusion that global and experiential measures of affect are quite similar in terms of their temporal stabilities and criterion-related validities, at least for the variables we considered.

Before discussing stability and criterion validity, it is important to note that global and experiential affective measures appeared to tap similar constructs in our study. Indeed, on a latent level, global and experiential positive affect correlated $r = 0.82$ with one another, and global and experiential negative affect were correlated $r = 0.72$. This suggests that the constructs captured by our global and experiential affect measures were quite similar to one another (albeit the correlation was not perfect). Notably, this represents a point of divergence from previous studies. For example, we previously found that the latent correlations between global and experiential affect were less than $r = 0.50$ (Hudson et al., 2017).

One potential explanation for this discrepancy could be design and sample differences between our two studies. For example, Hudson and colleagues (2017) studied a German sample three times across two years, and their measures of global and experiential affect contained different emotions (e.g., their measure of global positive affect consisted of a single emotion—happiness—whereas the experiential measure of positive affect contained multiple emotions). In contrast, in the present study, we examined a sample from Michigan across one month, and our global and experiential measures contained the same emotions. Any of these differences—or others (including sampling error)—may explain why we found that global and experiential measures are very similar to one another, whereas previous research has found greater divergence in these measures.

4.1. Stability in well-being

We examined the temporal stability in global self-report and experiential well-being over one month. Our findings indicated that global self-report and aggregated daily experiential measures of affect were remarkably similar in their test–retest stabilities. Approximately 75% of the variance in global self-reported positive and negative affect, as well as experiential positive affect was attributable to constant, trait-like latent variables. This translates into an average weighted test–retest correlation of $r = 0.75$. Experiential negative affect was relatively less stable than these other three variables with only approximately 60% of its variance attributable to a trait-like latent variable (which translates into an average weighted test–retest correlation of $r = 0.60$). Nonetheless, this is still a relatively large degree of stability. Thus, at least across one month, we found little evidence for differential stability in global versus experiential well-being.

It is important to note, however, that the observed stabilities in our study—including the percent variance explained by trait-like forces—represent only the portion of variance in each well-being variable that was constant *across the study's duration* (i.e., one month) (Anusic et al., 2012). Indeed, test–retest stabilities tend to decline over increasingly long test–retest intervals (Fraley et al., 2011; Fraley & Roberts, 2005). This occurs because situational forces, although not especially strong at any given occasion, can accumulate over time. Likewise, developmental influences that exhibit *autoregressive* properties and thus carry-over from one wave can also be present. These processes reflect slow changes that are hard to distinguish from trait variance in short-term studies. That is, autoregressive variance that has not fully decayed across the study's duration will be captured by test–retest correlations or state-trait models as unchanging trait variance (Anusic et al., 2012). Thus, long-term longitudinal studies are needed to completely distinguish between completely stable factors, slow-changing developmental or autoregressive trait factors (e.g., Kenny & Zautra, 2001), and state factors.

All told, long-term studies with multiple assessments allow researchers to obtain multiple estimates of test–retest stability over varying intervals and to identify the full range of possibilities for understanding stability and change (Fraley et al., 2011; Fraley & Roberts, 2005). The current study provides only one data point on this important issue. Nonetheless, meta-analyses already suggest that, in durations of less than one year, the stability of global well-being is approximately $r =$ 0.70—and that its stability asymptotically approaches approximately r = 0.20–35 over indefinitely long periods of time (Anusic & Schimmack, 2016; Schimmack & Oishi, 2005). Our finding that global well-being has one-month test–retest stabilities exceeding $r = 0.70$ is well aligned with these meta-analyses.

In contrast, fewer studies have examined temporal stability in experiential well-being as measured via DRM. Prior studies suggest that the stabilities for experiential well-being are approximately $r = 0.65$, 0.50, 0.40, and 0.35 over 2 weeks, 4 weeks, 1 year, and 2 years, respectively (Hudson et al., 2017, 2020; Krueger & Schkade, 2008). We found that the one-month stabilities for positive and negative affect were $r = 0.71$ and $r = 0.57$, respectively. These average to approximately $r = 0.65$—and thus mostly align with prior findings. Notably, these estimates of stability may have been inflated (as compared to global ratings) due to the fact that people rated their affect across multiple episodes. Nevertheless, these findings mirror those of Hudson and colleagues (2017)—who also found that, over a period of two years, global and experiential affect exhibited similar stability, with the exception that experiential negative affect was much less stable than any other type of well-being. Collectively, these findings may indicate that experiences of negative affect are both somewhat relatively rare (see descriptive statistics in Tables 1-2) and situationally-driven—whereas global patterns of negative affect may be more attributable to trait-like forces, such as individuals' levels of emotional stability.

Importantly, it is likely that our study underestimated the true stability in the well-being measures across time. Namely, most of our well-being scales contained two or fewer items (and one scale contained only a single item). As a consequence, we were unable to use latent variables to separate measurement error from state-level/contextual variation in well-being across all measures. Thus, our estimates of the portion of state variance in each measure are likely inflated by the presence of measurement error. Future research should employ affective measures with at least three items to be able to explicitly model measurement error and obtain more accurate estimates of the trait and state variance in each well-being measure.

4.2. Criterion validity of well-being

In addition to identifying patterns of stability overtime, we also examined the criterion validity of both global self-report and experiential measures of well-being. In our study, global and experiential affective measures exhibited very similar patterns of correlations with theoretically important criterion variables (cf. Diener & Tay, 2014). Both were related to personality traits, health, and socioeconomic status in theoretically predictable ways. Moreover, the sizes of the correlations were similar across experiential and global measures. This suggests the global self-reports and DRM-based experiential reports aggregated across up to three days have similar patterns of correlation with the external variables we investigated.

Our combined findings regarding stability and the criterion validity of experiential and global well-being measures suggest that experiential and global measures may function similarly to one another. Thus, our study indicates that DRM experiential measures are not necessarily superior to global measures (cf. Kahneman, 1999; Schwarz & Strack, 1999). Rather, both types of measures may provide similar information and might be equally valid. These findings should be replicated—but raise questions as to whether DRM experiential measures are necessarily desirable replacements for global self-reports, especially when considering their greater administration costs (e.g., the DRM can take upwards of an hour to complete; Kahneman et al., 2004).

Of course, it is possible that aggregating even additional reports of experiential emotional information could result in more a stable, accurate, and valid measure of participants' well-being—and potentially greater prediction of theoretically important criteria (Diener & Larsen, 1984). Such designs, however, may be infeasible—especially in large-scale survey work addressing multiple research aims. Moreover, the same logic may apply to help bolster the predictive validity of global measures. If global measures are truly contaminated by mood effects (Schwarz et al., 1987; Schwarz & Strack, 1999), it seems that collecting multiple measures of global well-being over short-periods of time and averaging them together should similarly cause any random, situational effects to mutually cancel, producing a more reliable estimate of well-being.

4.3. Consequences of weighting and sampling DRM episodes

As a final ancillary goal, we examined the psychometric consequences of two common methodological choices entailed in the DRM: (1) using duration-weighted versus unweighted affective composites, and (2) using abbreviated versions of the DRM in which participants rate affective experiences across only a subsample of reported episodes.

With respect to the former, scholars are divided regarding the optimal way to weight affective ratings in DRM episodes when forming composites. On one hand, some have argued that an individual's "objective" well-being should be the summation of his/her moment-by-moment experiences—and thus affective ratings in the DRM should be weighted by episode duration when composites are formed (i.e., reported emotions from longer episodes should "count more" in the overall composites; Kahneman, 1999; Kahneman et al., 2004). In contrast, other have argued that it may be more appropriate to used unweighted affective composites when analyzing data from the DRM (i.e., affective ratings from each episode are weighted equally irrespective of episode duration; Diener & Tay, 2014). This argument is based on the ideas that time may subjectively seem to pass more quickly or slowly for individuals during different episodes based on the activities in which they are engaged—and moreover, the psychological importance of any given moment may vary independently of real and/or perceived passage of time (Csikszentmihalyi, 1990). For example, an individual might perceive the holistic experience of riding a rollercoaster to be extremely positive, despite entailing upwards of 30–60 min of neutral affect while queuing, followed by perhaps only 90 s of intense positive affect while riding. Thus, the events that individuals freely choose to separate into self-defined episodes while completing the DRM may give better clues as to their psychological importance than the literal time devoted to those events.

Nevertheless, our findings suggested that the method used to weight DRM episodes when forming affective composites was largely immaterial. In terms of reliability, duration-weighted composites of experiential positive affect were slightly less stable than unweighted composites—but this was not true for experiential negative affect. In contrast, duration-weighted and unweighted affective composites were nearly identical in terms of their criterion validities. Thus, at least in our study, different choices regarding weighting of DRM episodes did not influence the psychometric properties of the measure.

Finally, with respect to using abbreviated versions of the DRM, the original "full" version of the DRM asks participants to list all activities in which they had engaged the prior day, and to rate their affective experiences in *all* episodes—which can take upwards of an hour (Kahneman et al., 2004). To increase the DRM's feasibility in common research contexts, researchers have created abbreviated DRM tasks in which participants list all activities in which they engaged, but only rate affective experiences for a few—usually three—randomly selected episodes (e.g., Anusic et al., 2017).

To emulate these shortened measures and explore their psychometric properties, we randomly selected three DRM episodes per participant per wave and used affective ratings in only these sampled episodes to form composites of experiential well-being. We repeated this random sampling procedure a total of 10 times—to create 10 "sampled affect" variables per participant per wave. Our results indicated that, on average across the ten random samples, as compared with the full DRM, experiential positive affect measured with the shortened DRM was slightly less stable across the span of one month. This indicates the presence of increased measurement error in shortened DRM assessments—which may somewhat reduce their ability to detect stable, trait-like individual differences in well-being as compared with the full-length DRM. Nevertheless, the patterns of correlations with external criteria were largely identical across the full and shortened versions of the DRM. This seems to indicate that, on average, randomly sampling DRM episodes should be expected to produce similar patterns of results to administering a full version of the DRM.

Indeed, this consistent with classical test theory (Lord & Novick, 1968). Namely, a person's affective rating in any given single DRM episode reflects their "true" level of trait affect plus random error (e.g., situation-specific variance, measurement error). Thus, the expected value of affect within a single DRM episode is equivalent to the expected value of affect in an aggregation of any number of DRM episodes.

That said, random errors only mutually cancel *in the long run* (i.e., with many observations). Thus, affect in any single DRM episode may provide a somewhat unreliable estimate of individuals' true trait levels of well-being. As an example of this principle, averaging across the 10 randomly sampled positive affect variables, the average correlation between sampled positive affect and agreeableness was $r = 0.07$ (compare to a correlation of $r = 0.08$ between for the full DRM positive affect composite). Yet in the individual random samples drawn, the correlation between agreeableness and sampled positive affect varied anywhere from $r = 0.02$ to $r = 0.11$. Thus, although studies that rely on a shorted version of the DRM in which affect is sampled from random episodes should not be expected to produce *systematically biased* correlations—the estimates from such studies may be more susceptible to random fluctuation than the estimates from studies that collect comprehensive DRM measures. In other words, aggregating across fewer DRM episodes may have the potential to increase Type I or Type II error rates—but not in any sort of systematic fashion.

That said, this conclusion must be tempered by the fact that our study did not experimentally randomize participants to complete the full version of the DRM versus an abbreviated version (instead, we randomly selected episodes reported by participants who had all completed the full DRM). Thus, there may be psychological processes that unfold while completing the full DRM versus an abbreviated version that might create differences between the measures. For example, when completing the full DRM, participants may experience fatigue or reactance, potentially compromising the quality of their data. Future research might consider experimentally manipulating whether participants complete either the full DRM versus an abbreviated version and examine whether there are any differences in the psychometric properties of the measures.

4.4. Conclusion

Scholars have argued that global measures of well-being may fundamentally lack validity and that experiential measures may represent a more optimal way of assessing well-being. Our study does not provide strong support for such a perspective. Instead, our findings suggest that global self-report and experiential measures of well-being from the DRM are more similar than different. Thus, well-being researchers may wish to increase their skepticism of sweeping critiques of self-report global measures. This study is, however, far from the last word on the topic—and future research should continue to directly compare and contrast global and experiential measures of well-being.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by a grant from the National Institutes of Health National Institute on Aging (#AG040715). We thank Ivana Anusic for assistance with data collection. Lucas and Donnellan collected the data. Hudson performed data analysis. All authors drafted and edited the manuscript.

References

Anusic, I., Lucas, R. E., & Donnellan, M. B. (2012). Dependability of personality, life satisfaction, and affect in short-term longitudinal data: Dependability in short-term studies. *Journal of Personality, 80*(1), 33–58. https://doi.org/10.1111/j.1467-6494.2011.00714.x

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. https://doi.org/10.1007/s11205-015-1172-6

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.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 1–12. https://doi.org/10.1038/nrn3475

Bylsma, L. M., Taylor-Clift, A., & Rottenberg, J. (2011). Emotional reactivity to daily events in major and minor depression. *Journal of Abnormal Psychology, 120*(1), 155–167. https://doi.org.proxy2.cl.msu.edu/10.1037/a0021662.

Cantril, H. (1967). *Human dimension experiences in policy research*. Rutgers University Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org.proxy2.cl.msu.edu/10.1037/h0040957.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optional experience*. Harper and Row.

Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95*, 542–575.

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*(4), 871.

Diener, E., Lucas, R., Schimmack, U., & Helliwell, J. (2009). Well-Being for Public Policy. *Oxford University Press*. https://doi.org/10.1093/acprof:oso/9780195334074.001.0001

Diener, E., & Tay, L. (2014). Review of the day reconstruction method (DRM). *Social Indicators Research, 116*(1), 255–267. https://doi.org/10.1007/s11205-013-0279-x

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*(2), 269–283. https://doi.org.proxy2.cl.msu.edu/10.1007/s11205-010-9578-7.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment, 18*(2), 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Dyrenforth, P. S., Kashy, D. A., Donnellan, M. B., & Lucas, R. E. (2010). Predicting relationship and life satisfaction from personality in nationally representative samples from three countries: The relative importance of actor, partner, and similarity effects. *Journal of Personality and Social Psychology, 99*(4), 690–702. https://doi.org/10.1037/a0020385

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*(7), 1097.

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review, 112*(1), 60–74. https://doi.org/10.1037/0033-295X.112.1.60

Fraley, R. C., Vicary, A. M., Brumbaugh, C. C., & Roisman, G. I. (2011). Patterns of stability in adult attachment: An empirical test of two models of continuity and change. *Journal of Personality and Social Psychology, 101*(5), 974–992.

Geiser, C., Keller, B. T., Lockhart, G., Eid, M., Cole, D. A., & Koch, T. (2015). Distinguishing state variability from trait change in longitudinal data: The role of measurement (non)invariance in latent state-trait analyses. *Behavior Research Methods, 47*(1), 172–203. https://doi.org/10.3758/s13428-014-0457-z

Heller, D., Watson, D., & Ilies, R. (2004). The role of person versus situation in life satisfaction: a critical examination. *Psychological Bulletin, 130*(4), 574–600. https://doi.org/10.1037/0033-2909.130.4.574

Helliwell, J. F. (2003). How's life? Combining individual and national variables to explain subjective well-being. *Economic Modelling, 20*(2), 331–360.

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

Hudson, N. W., Lucas, R. E., & Donnellan, M. B. (2017). 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., & Kushlev, K. (2016). Income reliably predicts sadness, but not happiness: a replication and extension of Kushlev, Dunn, & Lucas (2015). *Social Psychological and Personality Science, 7*, 828–836.

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

Kahneman, D., & Deaton, A. (2010). High income improves evaluations of life but not emotional well-being. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 16489–16493.

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*, 1776–1780.

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology, 63*(1), 52.

Kim-Prieto, C., Diener, E., Tamir, M., Scollon, C., & Diener, M. (2005). Integrating the diverse definitions of happiness: A time-sequential framework of subjective well-being. *Journal of Happiness Studies, 6*(3), 261–300. https://doi.org/10.1007/s10902-005-7226-8

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Welsley Publishing Company.

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

Lucas, R. E., & Dyrenforth, P. S. (2006). Does the existence of social relationships matter for subjective well-being? In K. D. Vohs, & E. J. Finkle (Eds.), *Self and relationships: Connecting intrapersonal and interpersonal processes* (pp. 245–273). Guilford Press.

Lucas, R. E., Tweten, C., Anusic, I., & Donnellan, M. B. (2016). *January). A direct comparison of the Experience Sampling Method (ESM) and the Day Reconstruction Method (DRM)* ([Poster].). San Diego, CA: Society for Personality and Social Psychology.

Macias, C., Gold, P. B., Öngür, D., Cohen, B. M., & Panch, T. (2015). Are Single-Item Global Ratings Useful for Assessing Health Status? *Journal of Clinical Psychology in Medical Settings, 22*(4), 251–264. https://doi.org/10.1007/s10880-015-9436-5

Nelson, S. K., Kushlev, K., English, T., Dunn, E. W., & Lyubomirsky, S. (2013). In defense of parenthood: Children are associated with more joy than misery. *Psychological Science, 24*(1), 3–10. https://doi.org/10.1177/0956797612447798

Okun, M. A., Stock, W. A., Haring, M. J., & Witter, R. A. (1984). The social activity/subjective well-being relation. *Research on Aging, 6*, 45–65.

Pierce, G. L. (2016). *Methodological report: Michigan State University State of the State Survey 1–70 Longitudinal Data*. Institute for Public Policy and Social Research: Michigan State University.

Pinquart, M., & Sorensen, S. (2000). Influences of socioeconomic status, social network, and competence on subjective well-being in later life: A meta-analysis. *Psychology and Aging, 15*, 187–224.

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

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

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

Robinson, M. D., & Clore, G. L. (2002b). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology, 83*(1), 198–215. https://doi.org/10.1037//0022-3514.83.1.198

Robinson, M. D., & Clore, G. L. (2007). Traits, states, and encoding speed: Support for a top-down view of neuroticism/state relations. *Journal of Personality, 75*(1), 95–120. https://doi.org/10.1111/j.1467-6494.2006.00434.x

Schimmack, U., & Oishi, S. (2005). The influence of chronically and temporarily accessible information on life satisfaction judgments. *Journal of Personality and Social Psychology, 89*(3), 395–406. https://doi.org/10.1037/0022-3514.89.3.395

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*(4), 582–593. https://doi.org/10.1037//0022-3514.82.4.582

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.

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*, 69–79.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological Momentary Assessment. *Annual Review of Clinical Psychology, 4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin, 134*(1), 138–161. https://doi.org/10.1037/0033-2909.134.1.138

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.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of breif measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070.

Wu, S., Wang, R., Zhao, Y., Ma, X., Wu, M., Yan, X., & He, J. (2013). The relationship between self-rated health and objective health status: A population-based study. *BMC Public Health, 13*, 320. https://doi.org/10.1186/1471-2458-13-320

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: Personality Processes and Individual Differences, 113*(6), 939–961. https://doi.org/10.1037/pspp0000115

Michigan State University Institute for Public Policy and Social Research. (2015). *State of the State Survey 69 [Data file and code book]*. http://ippsr.msu.edu/soss/.