HHS Public Access

Author manuscript

Emotion. Author manuscript; available in PMC 2019 October 01.

Published in final edited form as:

Emotion. 2018 October ; 18(7): 1024–1031. doi:10.1037/emo0000357.

The Short-Term Stability of Life Satisfaction Judgments

Richard E. Lucas[1], **Vicki A. Freedman**[2], and **Jennifer C. Cornman**[3]

[1]Department of Psychology, Michigan State University

[2]Institute for Social Research, University of Michigan

[3]Jennifer Cornman Consulting

Abstract

Life satisfaction judgments are thought to reflect people's overall evaluation of the quality of their lives as a whole. Because the circumstances of these lives typically do not change very quickly, life satisfaction judgments should be relatively stable over time. However, some evidence suggests that these judgments can be easily manipulated, which leads to low stability even over very short intervals. The current study uses a unique data set that includes multiple assessments of life satisfaction over both long (up to four years) and short (over the course of a single interview) intervals to assess whether information that is made salient during the course of an interview affects life satisfaction judgments at the end of the interview. Results suggest that this intervening information has only small effects on the final judgment and that placement within an interview has little influence on the judgment that people provide.

Keywords

Subjective well-being; life satisfaction; stability; measurement

Subjective well-being [SWB] is an overall evaluation of the quality of a person's life as a whole (Diener, Suh, Lucas, & Smith, 1999). SWB is usually assessed either by tracking people's emotional reactions to the conditions of their lives, or by explicitly asking respondents to reflect on their lives and to derive a global judgment of life satisfaction. These latter measures of life satisfaction, which can be obtained quickly and easily, have been used frequently in both applied and theoretical research in the social sciences. Their efficiency makes them suitable for a broad range of assessment settings, from experimental studies, to large-scale population-based surveys, to multi-wave panel studies spanning many years. Yet, concerns about their reliability and validity remain. The goal of this paper is to use a unique data set that includes multiple assessments of life satisfaction over both long (up to four years) and short (over the course of a single interview) intervals to address ongoing debates about the processes underlying these judgments, debates that have implications for the reliability, validity, and utility of these measures.

Constructing Life Satisfaction Judgments

When judging their life satisfaction, respondents presumably consider the objective conditions in their lives–perhaps in relation to some subjective standard—and then aggregate across those conditions to arrive at an overall judgment. Because the objective conditions themselves (conditions such as one's income, employment status, marital status, and health) are relatively stable over time, well-being judgments, too, should be somewhat stable, at least across relatively short time intervals. Thus, an important issue when assessing the validity of life satisfaction measures concerns the rank-order stability of these measures over time (for a broad review of issues related to stability, see Sheldon & Lucas, 2014)

Typically, studies that have examined the rank-order stability of life satisfaction judgments have done so over relatively long periods of time (i.e., months or years; see Eid & Diener, 2004; Lucas & Donnellan, 2012; Lucas, Diener, & Suh, 1996; Schimmack & Oishi, 2005). These studies consistently show that life satisfaction judgments exhibit reasonably strong stability. For instance, Schimmack and Oishi (2005) conducted a meta-analysis of stability coefficients and showed that one-year test-retest correlations hover around .50 for single-item measures and as high as .70 for multiple-item measures. They also showed that these correlations decline with increasing intervals and then approach an asymptote beyond which increasing intervals are not associated with further declines in stability. These patterns have also been confirmed in large-scale panel studies with tens of thousands of participants who have been studied for over 25 years (Lucas & Donnellan, 2012).

It is important to acknowledge, however, that less-than-perfect stability is likely due to some combination of true change in satisfaction, random measurement error, and—more problematically—systematic distortions. Indeed, in one frequently cited critique of global life satisfaction measures, Schwarz and Strack (1999) argued that the process that people use to derive life satisfaction judgments can lead to unstable responses. Specifically, they suggested that people do not conduct a thorough search of their memory for information that would be relevant for life satisfaction judgments. Instead, they rely on a variety of short-cuts and heuristics. For instance, in one famous study, Strack, Martin, and Schwarz (1988) manipulated the information that was salient for a group of college students by experimentally manipulating the order of two questions, one about dating frequency and one about satisfaction with life. When the dating frequency question was asked first, it was strongly correlated with life satisfaction; when the satisfaction question was asked first, the two measures were uncorrelated (though see Schimmack & Oishi, 2005, for contradictory evidence). In a separate study, Schwarz and Clore (1983) found evidence that college students relied on their mood when making life satisfaction judgments. They showed, for instance, that respondents who reported their life satisfaction on a pleasant sunny day reported higher scores than those who provided responses on cold, rainy days (though see Lucas & Lawless, 2013, Yap et al. (2016), for evidence that these effects are not robust). Importantly, the reduced rank-order stability that result would be a sign of weakened validity of SWB measures, as they reflect systematic distortions rather than the effects of random measurement error.

In developing the argument for their broader critique, Schwarz and Strack (1999) relied on the idea that short-term stability of life satisfaction measures is weak. Indeed, in introducing their critique, they stated that "measures of SWB have low test-retest reliabilities, usually hovering around .40, and not exceeding .60 when the same question is asked twice during the same one-hour interview" (p. 62). Schwarz and Strack (1999), like the authors they cited in support of this statement, interpreted this correlation as being disappointingly small. Ultimately, Schwarz and Strack concluded that this instability over short periods of time, when combined with experimental evidence that context matters, might mean that "life satisfaction judgments seem too context-dependent to provide reliable information about a population's well being" (p. 80). Thus, evidence about short-term stability of life satisfaction measures is important not only when focusing specifically on the reliability of these measures, but also when assessing the plausibility of process models that underlie these judgments and the validity and utility of the measures themselves.

The suggestion that survey content can strongly affect satisfaction judgments is related to a broader practical question about how life satisfaction—and other related constructs— should be assessed. If judgments like life satisfaction are strongly affected by random and idiosyncratically salient contextual information, then the placement of the question in a broader survey matters (see Deaton, 2012, for an example). On the one hand, researchers may wish to place life satisfaction questions near the beginning of a survey so as not to influence participants' responses by making information that would otherwise not come to mind become salient for participants. On the other hand, if life satisfaction measures are placed at the beginning of the survey, then the survey designers have less control over the information that is salient, and participants' responses could be influenced by a variety of unknown random or systematic factors that reduce the reliability and validity of the responses that are provided. This concern would argue for placement at the end of the survey, after similar information was made salient to all participants. At this point, there is little empirical evidence available about how the positioning of life satisfaction questions affects their reliability and validity.

In this study, we take advantage of a unique dataset that includes multiple measures of life satisfaction assessed multiple times over a period of years, along with two life satisfaction measures assessed at the beginning and end of a supplemental interview designed to assess disability and time use. By modeling a stable, trait-like component from the life satisfaction measures assessed across multiple years, it is possible to obtain a reasonable estimate of one's stable level of life satisfaction. It is then possible to determine which of the two measures that were assessed over a very short interval map mostly closely on to the stable level estimated from multiple years' worth of data. In addition, it will be possible to use the responses to the survey itself to determine whether life satisfaction questions presented at the end of a survey are more strongly related to the information that is made salient during that interview than are measures assessed at the beginning of the survey. The results we obtain will have practical implications for researchers who design surveys using life satisfaction measures. More importantly, however, these results will inform basic theories about reliability and stability of life satisfaction, along with the processes that underlie the judgments themselves.

Sample and Methods

We rely on the 2009–2013 waves of the U.S. Panel Study of Income Dynamics (PSID) and the 2013 Disability and Use of Time (DUST) supplement to the PSID. The PSID is the world's longest running national observational household panel study. Begun in 1968 with approximately 5,000 families, interviews have been conducted annually through 1997 and every two years by telephone since then with one household informant. When adult children leave home and form their own economically independent households, they become eligible for the study; consequently the sample grows naturally. Approximately 8,600, 8,900, and 9,100 interviews were conducted in 2009, 2011, and 2013, respectively; reinterview response rates for those years ranged from 91%–94%. In 2013, a supplemental study of adults ages 60 and older in the PSID and their spouses or partners was conducted by telephone to obtain detailed information on disability, use of time, and experienced wellbeing. 1,776 respondents participated in all (72% of eligible households completed at least one interview). Detailed information about all measures used in the PSID and DUST supplement are available at: https://psidonline.isr.umich.edu/guide/documents.aspx

The main PSID interview consists of questions about a variety of topics including income, program participation, employment, housing, expenditures, health, wealth, family composition, and education. In addition, since 2009, at the beginning of the interview, the respondent is asked about their life satisfaction. They are asked "Please think about your life as a whole. How satisfied are you with it? Are you completely satisfied, very satisfied, somewhat satisfied, not very satisfied, or not at all satisfied?" Over 6,900 respondents completed the life satisfaction measures in each of the three waves of the PSID.

The DUST supplement also includes life satisfaction questions, though the response scale differs from that used in the core study. The first question asks, "Please use a scale from 0 to 6, where 0 means not at all satisfied and 6 means very satisfied. Taking all things together, how satisfied are you with your life these days?" Respondents were then asked a series of additional questions about their physical impairments and limitations; their personality, self-efficacy, and spirituality; their memory; and their partner and family relationships. Next, respondents were asked to complete a time diary covering the prior 24 hour period (including what they were doing, how long it took, where they were and who they were with). Three activities were then selected for more detailed assessment of emotions ("calm", "happy", "frustrated", "worried", "sad", "tired", "pain") experienced during those activities. Final sections focused on caregiving, division of labor within the household, and participation in social and productive activities. At the end of the interview, respondents were asked, "Now that you have had a chance to think about how you spend your time, I'd like to ask you one final question. For this question, please use a scale from 0 – 6, where 0 means not at all satisfied and 6 means very satisfied. Taking all things together, how satisfied are you with your life these days?" The questions included in our analysis reflect the entire range of questions included in the DUST survey, though some specific questions that should be irrelevant for life satisfaction judgments (e.g., who, specifically, a person was with during specific activities) were not included. Full measures with exact question wording are presented in the online documentation.

Results

Correlations across the three waves of the PSID are reported in left panel of Table 1. As can be seen, these correlations ranged from .43 to .49, showing that life satisfaction measures were moderately stable even over relatively long periods of time. These correlations were also similar in size to those from other panel studies that included similar measures (Lucas & Donnellan, 2012). For instance, according to the figure presented on p. 297, Schimmack and Oishi (2005)'s meta-analytic estimates would predict two- and four-year stabilities of approximately .47 and .42. The estimates from the PSID were quite close to these predicted values. Importantly, the correlations in this study only decayed slightly with increasing length of interval. To see this, one can compare the two-year stabilities to the four-year stability coefficient. This pattern suggests that it is reasonable to model a stable latent trait from the three indicators, as opposed to using a latent autoregressive model (see Anusic, Lucas, & Donnellan, 2012, for a discussion of these models).

Next, we examined the associations between these three measures in the smaller subsample of respondents who also participated in the 2013 DUST. These correlations are reported in the right panel of Table 1. These correlations were quite similar in size to those from the broader sample, and the pattern also suggests a strong stable-trait component, rather than the clear decay associated with a strong autoregressive pattern.

The fourth and fifth lines of Table 1 also show the correlations between the pre- and post-DUST life satisfaction measures and each of the PSID measures of life satisfaction. The first thing to note is that the 0.62 correlation between these two measures was consistent with estimates provided by Schwarz and Strack (1999). In addition, the correlations with the measures from the PSID were comparable in size to the correlations among the PSID measures. As might be expected, the correlations between the DUST measures (which were assessed in 2013) and the 2013 PSID measure were slightly higher than the correlations with the 2011 and 2009 PSID measures, but the differences were not especially large. Most importantly, the correlations with the pre-DUST measure were very close in size to the correlations with the post-DUST measure, which provides initial evidence that the position of the item in the survey had little effect on the validity of the measure, when prior assessments of life satisfaction were used as a validity criterion.

Modeling a Stable Trait

The next goal of our analyses was to model a latent stable trait from the three PSID measures to determine the amount of variance in each assessment that is due to this stable trait. It is then possible to add the DUST measures to the model to see whether the stable component contributes more strongly to the pre-DUST measures or to the post-DUST measures. We first estimated a simple stable-trait model with a single latent trait with each wave of the PSID as indicators. The factor loadings were constrained to 1 for all indicators, a decision that reflects the assumption that the stable trait affects each measure in similar ways. With large samples, even very subtle sources of model misfit can lead to significant χ^2 values. Thus, model fit was evaluated using standard criteria for alternative fit indexes, such as CFI, RMSEA, and SRMR. Specifically, models with CFI values above .95, and

RMSEA and SRMR values below .05 generally indicate acceptable model fit (Hu & Bentler, 1999).

The model fit well (χ^2 = 5.40, df = 2, p = 0.067, CFI = 1.00, RMSEA = 0.04, SRMR = 0.03). Because all loadings were constrained to a value of 1, the only estimates of interest are the variances for the stable trait and the three indicators. These variances are presented in Table 2. We were most interested in the percent of variance in each indicator that can be accounted for by the stable trait. These values can be calculated by taking ratio of stable trait variance to total variance in the indicator, where total variance is the sum of the trait variance and the residual variance. According to these results, the stable trait component accounted for 46.64%, 50.29%, and 52.96% of the variance in the 2009, 2011, and 2013 measures, respectively. The slight increase in variance accounted for by the stable trait results from the fact that the life satisfaction measures were increasingly stable over time in this sample.

Next, we tested a similar model where the three PSID measures served as indicators of the stable latent trait (with all three loadings again constrained to 1) and the two DUST measures were predicted from this latent trait. The paths from the latent trait to these observed measures were estimated, and the covariances among the three measures assessed in 2013 were estimated (to account for any occasion-specific variance)[1]. This model also fit well (χ^2 = 15.32, df = 4, p = 0.004, CFI = 0.99, RMSEA = 0.05, SRMR = 0.02). The estimated unstandardized path coefficients from the latent trait to the two DUST measures were strong and significant (path to the pre-DUST measure: 1.04 SE = 0.06, z = 18.42, p < 0.001; path to the post-DUST measure: 1.13 SE = 0.06, z = 18.24, p < 0.001). The correlation between the residual variances for the two DUST measures was also relatively strong: r = 0.45, z = 10.05, p < 0.001. Although the correlations between each of these measures and the 2013 PSID life satisfaction measure were significant, they were small in size: r = 0.09, z = 2.07, p = 0.038 and r = 0.09, z = 2.04, p = 0.042 for the pre- and post-DUST measures, respectively.

Variance estimates for this modified model are presented in Table 3. Again, it is possible to use these estimates to calculate the amount of variance in the pre- and post-DUST measures that can be accounted for by the stable trait component (though the trait variance in Table 3 must be multiplied by the square of the loading reported in the previous paragraph, given that these loadings are not constrained to 1). Based on these estimates, the stable trait component accounted for 35.56% and 35.07% of the variance in the pre- and post-DUST measures of life satisfaction, respectively. These values were slightly lower than the percent of variance accounted for by the stable trait in each of the three PSID assessments (values that tend to fall closer to 50%). This could be due to the fact that the DUST measures use a different response scale and the stable-trait component included method variance that was shared across the three PSID measures. Alternatively, there may be specific contextual factors that influence measures assessed in a single occasion. It is important to note,

[1]This model is mostly equivalent to one where the two DUST measures are treated as additional indicators of the latent trait. The only difference is that the loadings for these observed variables are not constrained to one, to allow for the possibility that the DUST measures will relate differently to the underlying latent trait.

however, that the stable latent trait accounted for as much variance in the pre-DUST measures as in the post-DUST measure, suggesting that they are equally good indicators of this stable latent trait.

Correlations With Predictors

The primary concern about the position of life satisfaction questions within a survey is that the content of the survey will influence respondents' satisfaction judgments. If so, correlations between survey content and life satisfaction judgments should be higher when the life satisfaction question is administered at the end of the survey than they are when the question is located at the beginning of the survey. In the DUST interview, approximately one hour's worth of questions were administered between the two life satisfaction measures, and it is not clear whether the associations would be expected to increase equally for all intervening questions. On the one hand, it is possible that content that is temporally closest to the final life satisfaction questionnaire would exhibit the largest differences in correlations, as that content would be most salient at the time of judgment. However, Schimmack and Oishi (2005) also suggested that context effects should theoretically be largest for content that is especially relevant for life satisfaction but that is not already *chronically* salient to respondents. In other words, Schimmack and Oishi suggested that content that is typically not on respondents' minds but that would be considered relevant for life satisfaction judgments once it was made salient would show the most pronounced context effects. The intervening content in the DUST survey covers a range of topics that have traditionally been linked with SWB (such as personality, social support, health, and daily activities; see Diener et al., 1999, for a review), but it is not clear how the degree of chronic salience for each predictor would interact with the relevance and temporal closeness to the final satisfaction judgment to influence potential context effects. Thus, we made no predictions about which correlations will differ most strongly across the two administrations of the life satisfaction question.

Tables 4 through 10 show the correlations between each of the two life satisfaction measures and the additional predictors we assessed, grouped by content. In addition, for each pair of correlations, we report a t-test for dependent correlations that tests whether the correlations in that row are significantly different from one another. Due to length considerations, we do not discuss individual associations. Instead, we briefly review the reasons why each set of constructs might be expected to correlate with life satisfaction and then focus on broad conclusions about the differences in correlations that can be drawn.

Table 4 shows the correlations between the two life satisfaction measures and measures of the Big Five personality traits, a measure of self-efficacy, and a measure of spirituality. Prior research shows that individual differences in personality (and to a lesser extent religiosity) are some of the strongest correlates of SWB (Diener et al., 1999). Thus, making these domains salient could potentially affect life satisfaction judgments. Table 5 shows the correlations between various domain satisfaction ratings and the two life satisfaction judgments. Domain satisfaction ratings like these are often used in studies examining context effects on life satisfaction judgments (Schimmack & Oishi, 2005), and thus may be expected to be especially likely to elicit changes in reports of life satisfaction. Table 6

reports the results for three relationship variables: a composite score for the quality of one's relationship with his or her romantic partner, a composite score for the quality of one's relationship with his or her family, and the respondent's report of the number of friends he or she has. As with personality traits and domain satisfaction ratings, relationship quality has often been linked with reports of SWB (Myers, 2000). Table 7 reports the correlations between a series of questions on disability and impairment and the two life satisfaction measures.

Tables 8, 9, and 10 report the correlations between life satisfaction and a series of variables assessed towards the end of the survey, including variables based on the time-use measure (which made up a large portion of the survey), as well as the very final questions asked during the interview. Specifically, Table 8 reports the correlations between the life satisfaction questions and the number of minutes spent yesterday in different types of activities (from the time-use survey), Table 9 reports the correlations between the life satisfaction questions and various affective measures of well-being derived from the time-use survey, and Table 10 reports the correlations between the life satisfaction measures and a series of questions about the number of days various activities were performed during the previous seven days. These final questions immediately preceded the post-DUST life satisfaction measure and thus may have especially strong effects on life satisfaction judgments if temporal closeness plays a role.

Three patterns can be seen in Tables 4 through 10. First, the differences in correlations between the predictors and the pre-DUST and post-DUST measures of life satisfaction were mostly nonsignificant. Specifically, only 13 out of 48 correlations were significantly different from zero (not adjusting for multiple comparisons). Second, those differences that were significant were quite small, with maximum differences in correlations of just .07. Finally, it is important to note that of those differences that were significant, most (11 out of 13) involved cases where the correlation with the post-DUST measure was larger than the correlation with the pre-DUST measure, a finding that is consistent with the idea that intervening content could affect life satisfaction judgments.

The correlations reported in Tables 4 through 10 suggest that individual predictors were at best only slightly more strongly correlated with the post-DUST life satisfaction measure than with the pre-DUST measure. However, if each predictor added a small amount of incremental variance to the post-DUST measure, this could result in a measure that has substantial amounts of reliable variance that is distinct from the reliable variance included in the pre-DUST measure or that would have been included had the additional life domains not been made salient. One way to test this possibility is to predict the post-DUST measure from the pre-DUST measure and then add the additional predictors to the model to see how much additional variance the combined set can explain.

It is important to remember, however, that the initial judgments are measured with error. Thus, additional predictors may be associated with post-DUST life satisfaction even after controlling for pre-DUST life satisfaction simply because the pre-DUST measure does not adequately control for the associations between the predictors and stable levels of life satisfaction (Westfall & Yarkoni, 2016). To address this concern, we can also reverse the

direction of the regression analysis, predicting the pre-DUST measure from the predictors after controlling for the post-DUST measure. Because any additional prediction cannot be due to the information being made salient (as the pre-DUST measure was assessed before the predictors), this analysis can serve as a comparison by which the prior results can be judged.

Some of the questions from Tables 4 through 10 were only asked of certain respondents (e.g., marital satisfaction was only obtained from married participants), which means that the sample size would be substantially reduced if all predictors were included in a single model. Therefore, we reduced the set of predictors in such a way as to maximize sample size, while still including as many predictors as possible from the previous analyses. Specifically, we combined the domain satisfaction scores from Table 5 into a single measure that averaged across all non-missing domain satisfaction scores. In addition, because the partner quality measure led to the most missing data, we excluded this measures from the final model.

Because the estimated regression coefficients for the individual predictors are not of interest, the details of these analyses are not reported here (though they are available on the Open Science Framework page associated with this project: https://osf.io/d8f6t). Instead, we focus on the amount of variance that was explained across the three models tested. The baseline model, predicting post-DUST life satisfaction from pre-DUST life satisfaction had an R^2 value of 0.39. The complete model, which included all the predictors from Tables 4 through 10 (with exceptions described above) had an R^2 of 0.50. Thus, including the predictors led to an increase of 11.30% explained variance. As noted above, however, this value does not directly reflect the influence of the predictors on the post-DUST judgment, as at least some of the explained variance results from the fact that the pre-DUST judgment is measured with error. To get a sense of how much of the additional explained variance is due to this fact, it is possible to reverse the prediction, explaining pre-DUST life satisfaction from post-DUST life satisfaction and the set of predictors. This model had an R^2 of 0.48, which means that the predictors accounted for an additional 8.40% of the variance in the pre-DUST measure, even though these predictors were assessed *after* the pre-DUST measure was administered. This suggests that making this information salient only led to about 2.90% additional reliable variance in the final life satisfaction measure.

Discussion

Well-being judgments are thought to reflect people's overall evaluation of the quality of their lives as a whole. Because the circumstances of people's lives do not typically change very quickly, judgments that are supposed to be influenced by these circumstances should also be relatively stable, at least over relatively short periods of time. If, however, respondents' judgments are typically influenced by irrelevant contextual factors (such as mood at the time of judgment or the specific information that was arbitrarily made salient), then stability would be reduced, and the reliability and validity of well-being measures could be called into question (Schwarz & Strack, 1999). Past research that has examined these questions has either focused exclusively on long-term stability coefficients (e.g., Lucas & Donnellan, 2012; Schimmack & Oishi, 2005), or it has relied on experimental studies in which manipulated scores are typically not compared to long-term levels (e.g., Schwarz & Clore,

1983; Strack et al., 1988). The goal of the current study was to use a unique dataset to simultaneously address both long- and short-term stability of life satisfaction measures.

Consistent with past research, our results showed that life satisfaction scores are reasonably stable over a period of two to four years, with approximately half the variance in these scores accounted for by a stable trait. More importantly, we showed that regardless of whether a life satisfaction measure is administered at the beginning of a survey (where scores could be affected by idiosyncratic influences that may differ across individuals) or the end of a survey (where scores could be affected systematically by the content of the survey), scores on this measure tap this stable latent trait equally well. Thus, even though stability of the single-item life satisfaction item from the beginning to the end of the survey was only $r = 0.62$ (a correlation that some have considered to be surprisingly weak), the associations with scores from prior years was similar across the two assessments.

Furthermore, when we explicitly tested whether the survey content that was made salient during the course of the interview was more strongly related to the post-interview life satisfaction scores than to the pre-interview scores, we found only slight differences. Although the correlations between the predictors and the scores from the second assessment were often greater than those with scores from the first assessment (and sometimes significantly so), the absolute size of these differences was quite small, typically lower than .07. Importantly, when all predictors were simultaneously entered into a regression analysis predicting the post-interview life satisfaction scores, these predictors only contributed a small amount of additional variance over the pre-interview scores. Thus, the final scores do not differ strongly and systematically from the scores at the beginning of the interview.

So why aren't responses to the two measures identical? Given the pattern of results described above, it seems that random measurement error may be the primary explanation. All psychological constructs are measured with error, and scores from single, self-report questionnaire items may be especially likely to contain substantial amounts of it. Indeed, different items from the same life-satisfaction scale typically only correlate around .50 with one another, even when these items are assessed at the same point in time (Schimmack & Oishi, 2005). People who respond to single-item life satisfaction measures may mis-hear the question, they may not accurately communicate their response, or the interviewer may incorrectly key in the response option that the respondent provides. In addition, because respondents must translate their internal judgment to a discrete rating on a 5-, 7-, or 11-point scale, any slight differences in the specific response option that is chosen can lead to discrepant responses that add to measurement error.

These results are important because they provide insight into the reasons behind the less-than-perfect short-term stability of life satisfaction measures. If this instability is simply due to measurement error, then it has few implications for research findings that use these measures. Correlations between well-being measures and other predictors and outcomes will be attenuated when substantial amounts of measurement error exists. However, this problem can be addressed by using larger sample sizes (which have greater power to detect attenuated effects), multiple-item scales that reduce measurement error, or latent-variable modeling strategies that correct for measurement error. In short, random measurement error is not

especially problematic, as solutions exist for dealing with it. In contrast, systematic distortions, like those described by Schwarz and Strack (1999) are more problematic. For instance, if making disability status salient had a substantial effect on life satisfaction scores, then researchers may draw different conclusions about the importance of disability status for quality of life depending on whether disability questions were administered before or after the satisfaction measure. The current study shows that correlations are quite similar regardless of whether the life satisfaction measure came before or after this additional content. In addition, the results of our study suggest that the short-term instability of life satisfaction judgments is more likely due to random than systematic errors.

Although the research reported in this paper focuses specifically on context effects in life satisfaction judgments, the results also have implications for a broader range of phenomena. For instance, estimating the reliability and validity of self-report measures is often a concern in research on emotion. Because emotions change relatively quickly, some traditional approaches to assessing reliability, such as examining test-retest correlations, can be problematic. Yet at the same time, because emotion reports are often obtained *in situ*, emotion researchers must often sample a broad range of emotions using a relatively small number of items. In such cases, items may not be expected to cohere, which is a problem if internal consistency is used to estimate reliability. Recently, emotion researchers have proposed combining long- and short-term measures as a way of providing estimates of reliability that are more appropriate for constructs like affect and emotion (e.g., Chmielewski & Watson, 2009; Chmielewski, Sala, Tang, & Baldwin, 2016). The results of the current study provide further evidence that comparing short- and long-term stability coefficients in a single study can be a fruitful way to examine the psychometric properties of constructs like emotion and well-being, constructs that are expected to change over time.

Conclusion

Subjective well-being measures, such as the single-item life satisfaction scale used in this study, are not perfectly stable. This is due, in part, to the fact that the construct itself likely changes over time. Research shows, however, that even over very short intervals, stability is not perfect, and the explanation for this short-term instability has implications for researchers' understanding of the psychometric properties of the measures. In addition, a better understanding of the processes that underlie instability helps clarify the processes by which people make well-being judgments.

In this study, we examined patterns of stability over both very short and very long intervals, with the aim of understanding why responses to these measures shift even over the course of a relatively short interview. Our analyses show that both the pre- and post-interview scores are related to a stable latent trait in similar ways, suggesting that the inclusion of survey content did not shift respondents' scores from what they would otherwise have said. In addition, the intervening content that was introduced in the survey did not seem to influence the final judgments, as responses to these questions were only slightly more strongly correlated with the post-interview scores as compared to the pre-interview scores. These results—which are consistent with past research, both on the stability of well-being measures and on the psychometric properties of single items—suggest that a parsimonious

explanation of the instability is that it is due more to measurement error than to systematic distortions. This is an important finding because issues involving measurement error are often easier to deal with than are systematic biases that result from the survey context.

Acknowledgments

This research was supported by funding from the National Institute on Aging (P01AG029409 and R01AG040715).

References

Anusic I, Lucas RE, Donnellan MB. Dependability of Personality, Life Satisfaction, and Affect in Short-Term Longitudinal Data. Journal of Personality. 2012; 80(1):33–58. DOI: 10.1111/j.1467-6494.2011.00714.x [PubMed: 21241303]

Chmielewski M, Watson D. What Is Being Assessed and Why It Matters: The Impact of Transient Error on Trait Research. Journal of Personality and Social Psychology. 2009; 97(1):186–202. DOI: 10.1037/a0015618 [PubMed: 19586248]

Chmielewski M, Sala M, Tang R, Baldwin A. Examining the construct validity of affective judgments of physical activity measures. Psychological Assessment. 2016; 28(9):1128–1141. doi:http://dx.doi.org.proxy2.cl.msu.edu/10.1037/pas0000322. [PubMed: 27537007]

Deaton A. The financial crisis and the well-being of Americans 2011 OEP Hicks Lecture. Oxford Economic Papers. 2012; 64(1):1–26. [PubMed: 22389532]

Diener E, Suh EM, Lucas RE, Smith HL. Subjective well-being: Three decades of progress. Psychological Bulletin. 1999; 125:276–302.

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

Hu L-tBentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 1999; 6(1):1–55. DOI: 10.1080/10705519909540118

Lucas RE, Donnellan MB. Estimating the Reliability of Single-Item Life Satisfaction Measures: Results from Four National Panel Studies. Social Indicators Research. 2012; 3:323–331.

Lucas RE, Lawless NM. Does life seem better on a sunny day? Examining the association between daily weather conditions and life satisfaction judgments. Journal of Personality and Social Psychology. 2013; 104(5):872–884. DOI: 10.1037/a0032124 [PubMed: 23607534]

Lucas RE, Diener E, Suh E. Discriminant validity of well-being measures. Journal of Personality and Social Psychology. 1996; 71(3):616–628. [PubMed: 8831165]

Myers DG. The funds, friends, and faith of happy people. American Psychologist. 2000; 55(1):56–67. [PubMed: 11392866]

Schimmack U, Oishi S. The Influence of Chronically and Temporarily Accessible Information on Life Satisfaction Judgments. Journal of Personality and Social Psychology. 2005; 89(3):395–406. DOI: 10.1037/0022-3514.89.3.395 [PubMed: 16248721]

Schwarz N, Clore GL. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. Journal of Personality and Social Psychology. 1983; 45(3):513–523. DOI: 10.1037/0022-3514.45.3.513

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

Sheldon KM, Lucas RE. Stability of happiness: Theories and evidence on whether happiness can change. London: Academic Press; 2014.

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

Westfall J, Yarkoni T. Statistically Controlling for Confounding Constructs Is Harder than You Think. PLOS ONE. 2016; 11(3):e0152719.doi: 10.1371/journal.pone.0152719 [PubMed: 27031707]

Yap SCY, Wortman J, Anusic I, Glenn S, Scherer LD, Donnellan MB, Lucas RE. The Effect of Mood on Judgments of Subjective Well-Being: Nine Tests of the Judgment Model. Journal of Personality and Social Psychology. 2016; doi: 10.1037/pspp0000115

Table 1

Correlations between life satisfaction measures in each of the PSID waves for those in the PSID (left panel), and those only in the DUST (right panel)

	2009	2011	2013	2009	2011	2013	Pre-DUST	Post-DUST
2009	1.00			1.00				
2011	0.47	1.00		0.45	1.00			
2013	0.43	0.49	1.00	0.49	0.54	1.00		
Pre-DUST				0.45	0.39	0.49	1.00	
Post-DUST				0.42	0.39	0.48	0.62	1.00
mean	3.73	3.81	3.81	3.84	3.87	3.92	5.01	5.01
sd	0.88	0.87	0.86	0.84	0.82	0.82	1.02	1.12

Note. Total N for left panel using listwise deletion = 6916, but correlations reported above are based on pairwise deletion. Average pairwise N = 7527. Total N for right panel using listwise deletion = 1129, but correlations reported above are based on pairwise deletion. Average pairwise N = 1150.

Table 2

Variance Estimates from Stable Trait Model Using PSID Data

	est	se	z	pvalue	ci.lower	ci.upper
2009 Life Satisfaction	0.39	0.02	18.36	< 0.001	0.35	0.43
2011 Life Satisfaction	0.34	0.02	17.48	< 0.001	0.30	0.38
2013 Life Satisfaction	0.30	0.02	16.75	< 0.001	0.27	0.34
Stable Trait	0.34	0.02	17.55	< 0.001	0.30	0.38

Note. N = 1128.

Table 3

Variance Estimates from Stable Trait Model With PSID and DUST Measures

	est	se	z	pvalue	ci.lower	ci.upper
2009 Life Satisfaction	0.37	0.02	18.49	< 0.001	0.33	0.41
2011 Life Satisfaction	0.35	0.02	18.17	< 0.001	0.31	0.39
2013 Life Satisfaction	0.31	0.02	16.77	< 0.001	0.27	0.34
Pre-DUST Life Satisfaction	0.66	0.04	18.32	< 0.001	0.59	0.74
Post-Dust Life Satisfaction	0.81	0.04	18.40	< 0.001	0.72	0.90
Stable Trait	0.34	0.02	17.52	< 0.001	0.30	0.38

Note. N = 1121.

Table 4

Correlations between individual difference measures and pre-DUST and post-DUST measures of life satisfaction.

	Pre-Dust	Post-Dust	t	p
Extraversion	0.135	0.136	−0.034	0.973
Agreeableness	0.053	0.115	−2.880	0.004
Conscientiousness	0.105	0.142	−1.723	0.085
Neuroticism	−0.277	−0.315	1.855	0.064
Openness	0.021	0.058	−1.711	0.087
Self-efficacy	0.284	0.331	−2.305	0.021
Spirituality	0.046	0.101	−2.578	0.010

Note. N = 1593. The t-test reported in Column 3 tests for differences between correlated correlations.

Table 5

Correlations between domain satisfaction scores and pre-DUST and post-DUST measures of life satisfaction.

	Pre-DUST	Post-DUST	t	p
Health	0.479	0.427	2.876	0.004
Memory	0.234	0.218	0.773	0.440
Financial	0.422	0.370	2.782	0.005
Job	0.426	0.422	0.104	0.917
Daily Activities	0.531	0.536	−0.261	0.794
Marriage	0.425	0.404	0.921	0.357
Relationship	0.445	0.464	−0.275	0.784

Note. N for Rows 1, 2, 3, and 5 = 1756; N for Row 4 = 642; N for Row 6 = 1188; N for Row 7 = 121. The t-test reported in Column 3 tests for differences between correlated correlations.

Table 6

Correlations between relationship variables and pre-DUST and post-DUST measures of life satisfaction.

	Pre-Dust	Post-Dust	t	p
Relationship Quality	0.372	0.377	−0.233	0.815
Family Quality	0.147	0.204	−2.782	0.005
Number of Friends	0.117	0.167	−2.433	0.015

Note. N for Row 1 = 1309; N for Row 2 = 1756; N for Row 3 = 1756. The t-test reported in Column 3 tests for differences between correlated correlations.

Table 7

Correlations between impairment variables and pre-DUST and post-DUST measures of life satisfaction.

	Pre-Dust	Post-Dust	t	p
Health Impairments	−0.267	−0.296	1.474	0.141
Activity Limitations	−0.210	−0.252	2.088	0.037
Memory Rating	−0.181	−0.230	2.454	0.014
Memory Change	−0.101	−0.127	0.748	0.455
Memory Aids	−0.071	−0.047	−0.681	0.496

Note. N = 1754. The t-test reported in Column 3 tests for differences between correlated correlations.

Table 8

Correlations between activity variables and pre-DUST and post-DUST measures of life satisfaction.

	Pre-Dust	Post-Dust	t	p
Sleep	−0.012	−0.003	−0.423	0.673
Grooming	0.001	−0.021	1.046	0.296
Travel With Other	0.020	0.034	−0.702	0.483
Other Travel	0.041	0.069	−0.833	0.405
Work for Pay	−0.056	−0.024	−0.920	0.358
Socializing	−0.048	−0.075	0.784	0.433
Chores/Errands	0.014	0.038	−0.718	0.473
Providing Care	0.088	0.108	−0.620	0.535
Other	0.005	−0.008	0.380	0.704

Note. N = 1740. The t-test reported in Column 3 tests for differences between correlated correlations.

Table 9

Correlations between experiential well-being variables and pre-DUST and post-DUST measures of life satisfaction.

	Pre-Dust	Post-Dust	t	p
Total Pleasant Minutes	0.190	0.255	−3.199	0.001
Average Happiness	0.343	0.395	−2.723	0.007
Average Calm	0.229	0.245	−0.789	0.430
Average Frustrated	−0.211	−0.233	1.118	0.264
Average Worried	−0.213	−0.248	1.715	0.086
Average Sad	−0.207	−0.253	2.283	0.023

Note. N = 1740. The t-test reported in Column 3 tests for differences between correlated correlations.

Table 10

Correlations between activity variables and pre-DUST and post-DUST measures of life satisfaction.

	Pre-Dust	Post-Dust	t	p
Worked	−0.005	−0.013	0.380	0.704
Volunteered	0.072	0.107	−1.633	0.103
Provided Care	−0.006	0.031	−1.742	0.082
Socialized	0.079	0.154	−3.529	< 0.001
Exercised	0.111	0.129	−0.863	0.388
Went Out for Enjoyment	0.124	0.149	−1.191	0.234
Did Laundry	−0.013	0.006	−0.872	0.384
Cleaned/Did Repairs	0.059	0.082	−1.044	0.297
Made Dinner	−0.014	0.024	−1.796	0.073
Handled Finances	−0.018	−0.028	0.489	0.625
Shopped	0.030	0.031	−0.018	0.986

Note. N = 1593. The t-test reported in Column 3 tests for differences between correlated correlations.