Dependability of Personality, Life Satisfaction, and Affect in Short-Term Longitudinal Data

Ivana Anusic, Richard E. Lucas, and M. Brent Donnellan
Michigan State University

ABSTRACT The consistency of individual differences across time has implications for theory building and clinical applications. Indeed, personality psychologists have long worked to place constructs on the continuum of consistency of more trait-like to more state-like constructs. Recently, Chmielewski and Watson (2009) highlighted the importance of dependability coefficients for interpreting the results of stability studies. These coefficients provide an estimate of how strongly short-term transient error affects retest correlations for a given measure. In this article, we use a modified version of Kenny and Zautra's (1995, 2001) STARTS model to estimate dependability of personality, life satisfaction, and affect in a 2-month longitudinal study of 8 waves. Results from 226 undergraduate students indicated that personality ratings were least influenced by transient state factors, whereas affect was most influenced. We discuss these findings in terms of their implications for the continuum of consistency and for the practical issue of selecting retest intervals for dependability analyses.

The long-term consistency of psychological constructs is a topic of enduring interest in psychology (e.g., Conley, 1984; Ferguson, 2010; Roberts & DelVecchio, 2000; Trzesniewski, Donnellan, & Robins, 2003). Some of the most basic questions that personality psychologists seek to answer concern the consistency and stability of the constructs they examine. For instance, one of the most enduring debates in the field concerns William James's (1890) proposal that adult personality is set like plaster by age 30 (e.g., Ferguson, 2010; Roberts & DelVecchio, 2000; Terracciano, McCrae, Brant, & Costa, 2005). Thus, knowing how stable personality constructs really are is a central goal for personality research.

Correspondence concerning this article should be addressed to Ivana Anusic, Department of Psychology, Michigan State University, Psychology Building, East Lansing, MI 48824. Email: anusiciv@msu.edu.

Journal of Personality 80:1, February 2012

DOI: 10.1111/j.1467-6494.2011.00714.x

Yet information about the consistency of individual differences is not purely of theoretical interest; this information also has applied value. To be sure, basic research on the stability of personality does not tell researchers what is ultimately possible with intervention, but information about stability may inform expectations about the efficacy of such interventions. It is plausible to assume that individual differences that are highly stable over time may be more difficult to change by therapeutic interventions than are less stable constructs (Costa & McCrae, 1986, 1997).

In short, research concerning the consistency of psychological attributes is important from both theoretical and applied perspectives. Unfortunately, however, the interpretation of consistency information can be challenging. Although it is tempting to simply examine a single test-retest correlation and then to interpret whether the correlation is large or small, examining these single correlations in isolation has the potential to provide an incomplete and perhaps misleading picture of the stability of a construct (Conley, 1984; Fraley & Roberts, 2005). The appropriate interpretation of any one test-retest correlation requires a deeper understanding of the pattern of changes in these stabilities over varying intervals. The specific issue that the current article addresses concerns the idea that the interpretation of long-term stability coefficients requires insights into the short-term stability of constructs (Chmielewski & Watson, 2009). To address this issue, we use a modified version of state-trait statistical models (e.g., Cole, 2006; Kenny & Zautra, 1995, 2001; Steyer, Schmitt, & Eid, 1999) to examine the pattern of stability coefficients over varying lengths of time. This allows us to better isolate sources of transitory influences on psychological measures, which, in turn, helps inform judgments about the stability of constructs. Moreover, the present investigation can also provide evidence that different constructs have different levels of consistency even in the short term.

The Continuum of Consistency

One important insight from decades of research on the stability of psychological attributes is that the degree of consistency varies across psychological characteristics. Some characteristics (e.g., life satisfaction) are less consistent and thus presumably more responsive to changing life circumstances than others (e.g., personality). Conley (1984) proposed a continuum of longitudinal consistency of psychological traits, ranging from most stable (i.e., intelligence) to least stable (i.e., measures of self-opinion; but see Trzesniewski et al., 2003). This continuum theoretically extends from constructs that are pure traits to those that are pure states. Constructs that are pure states are short-lived and fluctuate from one moment to the next. Factors that influence these constructs are moment specific, and thus it is impossible to predict one's future standing on the construct from the present standing. In contrast, constructs that are pure traits are immutable even over long periods of time. In statistical terms, an individual's standing on the construct at any given moment will perfectly forecast her or his standing in the future.

Although the distinction between pure traits and pure states is conceptually simple, few, if any, psychological constructs can be found at these two extremes. Rather, psychological constructs involve both state-like and trait-like elements (Hertzog & Nesselroade, 1987; Nesselroade, 1991). Kenny and Zautra (2001) proposed a general framework for conceptualizing the psychological elements that contribute to observed variance in psychological constructs. According to their Stable Trait, Autoregressive Trait, State (STARTS) model, an individual's standing on a psychological construct is influenced by three latent variables—a completely stable element (i.e., the stable trait, or ST), a relatively enduring element that persists from one occasion to the next (i.e., the autoregressive trait, or ART), and completely transitory influences or state factors (i.e., the state, or S). The stable trait captures immutable attributes that generate completely stable individual differences in rank ordering. The autoregressive traits reflect elements of psychological attributes that ultimately change over time but nonetheless persist over shorter intervals. Last, state influences reflect the influence of transitory factors that are unique to a single occasion in time. Accordingly, the STARTS model provides a compelling general framework for conceptualizing the factors that contribute to stability and change in psychological attributes. The STARTS model also offers a specific analytic tool for decomposing variance in psychological attributes into these three attributes so they can be placed on the trait-state continuum.

As it stands, empirical research on the consistency of individual differences has not focused on STARTS model decompositions. Instead, much of the existing research has focused on documenting the level of observed consistency across time intervals of varying lengths using retest correlations (e.g., Conley, 1984; Low, Yoon, Roberts, & Rounds, 2005; Roberts & DelVecchio, 2000; Trzesniewski et al., 2003). Across psychological constructs such as personality, intelligence, self-esteem, and vocational interests, the patterning of retest correlations with respect to the length of the retest interval is similar and is characterized by three features.

First, retest correlations are never 1.0, even across very short retest intervals. This observation is consistent both with the proposition that some variation in measures is only temporary and with the proposition that measurement errors attenuate retest correlations. Conventionally, however, this observation is usually interpreted as providing evidence that state factors are a relevant influence on psychological measurement. The second feature of the observed pattern of retest correlations is that correlations decrease with increasing retest interval. This pattern provides support for the evidence of autoregressive trait factors. Indeed, a simplex-like pattern is consistent with the idea that accumulating changes makes it more difficult to predict one's standing on a construct over longer periods of time. The third feature of the observed pattern of retest correlations is that retest correlations do not reach zero even across intervals that span several decades. Instead, they approach an asymptote, whose value varies across constructs. This trend likely reflects the existence of stable traits—an immutable core that generates consistent individual differences across time.[1]

In short, the existing empirical database on the consistency of individual differences provides broad (albeit indirect) support for the STARTS model framework for conceptualizing psychological attributes. It is important to emphasize, however, that observed stability coefficients will carry different meanings depending on the retest interval over which they are assessed. Over long enough intervals (say, 15 or 20 years), these coefficients will largely reflect the influence of stable traits, whereas over very short intervals they reflect a mixture of stable trait and autoregressive trait influences.

It is also important to note that the interpretation of longer-term stability coefficients is affected by short-term stability coefficients.

1. In this article, we are concerned with rank-order stability, which is often part of the defining features of psychological traits. Rank-order stability is independent of mean-level stability. For example, although conscientiousness may increase with age, the relative ordering of individuals may still be preserved. Those who were more conscientious relative to their peers as adolescents may also be more conscientious relative to their peers as adults.

Indeed, Chmielewski and Watson (2009) have drawn attention to the importance of short-term stability coefficients for interpreting the magnitude of stability estimates. Short-term stability coefficients, which are often called dependability coefficients, assess the degree of transitory error in a measure. These forms of transitory error are not reflected in indexes such as coefficient alpha, but they are unique to a single occasion of measurement. Information about dependability should affect the interpretation of longer-term stability coefficients. For example, one might interpret a 5-year stability coefficient of .65 for a measure of extraversion differently depending on whether the short-term, 1-week stability of the measure (its dependability) was .70 versus .90. In the former case, almost all of the variance that persists over a 1-week period also persists for 5 years; in the latter case, it appears that more real change in the construct occurs.

The crucial issue for dependability coefficients is finding the right interval of time for estimating such effects. Conceptually speaking, time intervals should be short enough to preclude the possibility of real change but long enough to rule out the possibility that participants simply recall answers from memory (Heise, 1969; Watson, 2004). Unfortunately, it can be difficult to translate this conceptual idea into actual research practice.

Separating Transient Error From True Change With Trait-State Statistical Models

Indeed, a troubling practical issue for dependability studies is that it is difficult to specify the length of short-term intervals that is required to properly estimate dependability coefficients. Such design considerations depend on the construct in question (i.e., is it more trait-like or state-like) as well as the characteristics of the sample used for such studies (e.g., their developmental stage; Chmielewski & Watson, 2009). Some constructs may exhibit small but real changes even over short time intervals, and these changes could accumulate over time to produce meaningful changes in individual differences over longer intervals. Such small changes, however, would be classified as transitory error if the retest interval used to establish measure dependability was too long for the construct in question. Thus, it is useful to separate small, accumulating changes in individual differences over time from true transitory influences to establish optimal retest intervals for estimates of dependability.

The approach we propose for addressing this issue is to use the STARTS framework to help separate meaningful short-term change from transitory factors. The starting point is an intensive repeated measures strategy that is based on the shortest interval of time that is practically feasible (in this case, 1 week). The idea is to collect repeated 1-week assessments over a reasonable period of time (e.g., 2 months) to generate the kind of data that can be used with the analytic STARTS model. This model is simply a statistical tool that can be used to decompose variance in repeated measures assessments into the three conceptual sources of variance we previously outlined. In its basic form, the model posits that an individual's score at any one assessment is partitioned into three sources of variance: the stable trait (ST), the autoregressive trait (ART), and the state (S) factors.

It is important to note that the interpretation of the estimates provided by a STARTS model analysis will depend on the length of the retest interval and the duration of the study. Recall that state variance is defined as variance that is specific to a single measurement occasion and does not carry over from one measurement occasion to the next one. The proper interpretation of this coefficient requires a consideration of the particular study. Thus, for example, in STARTS analysis based on yearly assessments of life satisfaction (e.g., Lucas & Donnellan, 2007), the state component will reflect influences on life satisfaction that are specific to a single year. An athlete with a broken leg may not be able to engage in valued activities (e.g., running) for a few months, and thus she may rate her life satisfaction lower at a given time point. This source of variance is unlikely to have an impact on life satisfaction a year later once the injury has fully healed. However, state factors provide an important way to estimate dependability when studies are based on much shorter intervals. In such a context, it might be reasonable to assume that *1—state variance* is an indicator of dependable variance in a given measure. A short-term interval of, say, a week or two is probably too short to observe much real change in many constructs of interest for personality psychology. Moreover, if the model is applied to latent variables based on multiple indicators rather than observed scores, the S component will be free of measurement errors of the sort captured by coefficient alpha and thus reveal true transient influences on constructs (e.g., different psychological states specific to each occasion). Such a model applied to short-term data may therefore give a relatively pure measure of dependability. If the model is not applied to latent variables at each occasion, then occasion-specific variance and random measurement error variance are contaminated.

The design of a study also influences the interpretation of the other two components of the STARTS model—the autoregressive and stable traits. The ART variance reflects any factors that exert their influences for a time period longer than the retest interval but shorter than the duration of the study. In contrast, the ST variance reveals any influences on the scores that last over the entire duration of the study. This distinction is important because the duration of the study dictates the respective interpretations of the ART and ST components. In order for the ST to really be reflective of the true trait (i.e., stable individual differences), studies should be long enough for the correlation of the autoregressive components at first and last assessment to reach zero. If any of the autoregressive variance from the first wave is still present at the last wave, the model will have difficulty distinguishing this variance from the stable trait variance that is equally present at all waves. For many constructs, a large part of the autoregressive variance from the first wave is still present 1 year later (Anusic & Schimmack, 2011); thus, the ideal situation for completely isolating ST and ART variance is to have a longitudinal design that spans several years. This type of design, however, is not maximally useful for determining dependability. Nonetheless, the logic of the STARTS model can be applied to short-term longitudinal data with some modifications to provide more precise estimates of dependability, as we demonstrate in this study.

The Present Study

In sum, studies of the long-term consistency of individual difference hold an important place in personality psychology because they help researchers understand the nature of the constructs in question. However, the appropriate interpretation of such studies requires an appreciation of the dependability of measures. To address this issue, the present study provides a novel application of the STARTS analytic model to a 2-month longitudinal study in which participants made weekly ratings of their personality, life satisfaction, and affect. Our purpose was twofold. First, we aimed to develop a method for estimating dependability of a construct that goes beyond simple retest correlations. For this purpose, we modified the STARTS model for use with short-term longitudinal data with short retest intervals. This method can be useful for establishing adequate retest intervals for dependability estimates for the three constructs. For example, a construct may have a substantial autoregressive component that contributes to small, but real, changes that can accumulate over the period of 2 months. In that case, a retest interval shorter than 2 months may be more appropriate for the construct in question.

Second, we wanted to compare the amounts of transitory influences in personality, life satisfaction, and affect even in the short term. Evidence of differences would illustrate how short-term but intensive longitudinal studies can be used to determine where constructs fall on the continuum of consistency (Conley, 1984). This information has important theoretical consequences, such that it might indicate which constructs are more affected by momentary influences. For example, we can evaluate whether life satisfaction is more influenced by momentary states than are personality traits. In addition, information about dependability of the constructs can be used to interpret the extent of long-term stability of the three constructs. All in all, our goal was to illustrate how the STARTS analytic model applied to short term but nonetheless intensive longitudinal studies can cast light on crucial substantive and methodological issues in personality. We anticipate that affect judgments will be most influenced by momentary states (e.g., mood effects), whereas personality judgments will be least affected by such short-lived influences. Thus, we predict that personality will have the least amount of state influence, followed by life satisfaction and then affect.

METHOD

Participants

Two hundred thirty-seven undergraduate psychology students (85% female) at Michigan State University participated in an eight-wave study of personality over time. Eleven of these participants had missing data for five or more waves and were dropped from the analyses. The final sample consisted of 226 participants (85% female) with a mean age of 18.8 years ($SD = 1.4$ years). At each wave, participants completed a set of questionnaires that assessed life satisfaction, personality, and affect. Retest intervals were 1 week in length, with the exception of the interval between Waves 6 and 7, which was 2 weeks long due to spring break. For the first wave of the study, participants came into the lab, where they completed the paper version of the questionnaire. For the remaining waves, participants completed the questionnaire online.

Measures

Life satisfaction. Life satisfaction was assessed with the Satisfaction With Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985). The SWLS is a five-item global life satisfaction scale that assesses cognitive judgments of subjective well-being. Participants rated statements such as "In most ways my life is close to my ideal" on a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*). For this analysis, we selected only the first three items because they have been shown to have better psychometric qualities than items 4 and 5.

Personality. Participants completed the 20-item Mini-IPIP scale (Donnellan, Oswald, Baird, & Lucas, 2006; see also Cooper, Smillie, & Corr, 2010), which is based on the International Personality Item Pool (IPIP). This scale assesses each of the Big Five personality dimensions with four items. Scale instructions asked participants to describe how accurately each of the 20 statements describe them in general (e.g., "Am the life of the party," "Have frequent mood swings"). Participants made their ratings on a 5-point Likert scale ranging from 1 = *very inaccurate* to 5 = *very accurate*.

Affect. Affect was measured with a short version of the Mood and Anxiety Symptoms Questionnaire (Mini-MASQ; Clark & Watson, 1995). The scale yields three subscales: General Distress (8 items; e.g., "Felt tense or 'high strung,' " "Felt worthless"), Anxious Arousal (10 items; e.g., "Was short of breath," "Felt dizzy or lightheaded"), and Anhedonic Depression (8 items; e.g., "Felt like nothing was very enjoyable," "Felt like I had a lot to look forward to" [reverse scored]). Participants were asked to report how much they had felt or experienced each of the symptoms during the past week on a 5-point Likert scale (1 = *not at all*, 5 = *extremely*). Inspection of individual items indicated that general distress and anxious arousal item scores were positively skewed. In order to make the item distributions closer to normal distribution, we transformed the general distress items by taking the natural log of the original scale items. Because the skew was greater for anxious arousal items, we transformed these items by taking the inverse of the original score. It is important to note that these transformations merely improved the fit of the models and did not substantially change the conclusions we draw from our analyses.

Modeling Approach

Creation of latent variables. Separate analyses were conducted for each class of construct (personality traits, life satisfaction, and affect). As the first step of our analysis, we created a latent variable at each wave. For the five personality dimensions and for the life satisfaction scale, the items themselves were used as indicators. The factor loading of each item was constrained to be equal across waves (e.g., loading of Item 1 at Wave 1 = loading of Item 1 at Wave 2). Because the number of items for each of the subscales of the Mini-MASQ was substantially larger (8 to 10), we first created item parcels (three per subscale) and then created latent variables from the parcels. We randomly assigned items into three parcels for each of the three subscales. For general distress and anhedonic depression, the first two parcels were defined by three items, and the third parcel was defined by two items. For anxious arousal, the first parcel was defined by four items, and the other two parcels by three items each. Parcels were computed as the average of their items when data for all items were available, or they were set as missing when the data for some or all of the items were missing.

Modification of the STARTS model. We initially tried to test the full STARTS model, which specified stable trait, autoregressive, and state influences on all constructs. As expected, this model had problems with identification for most constructs. When the model could be identified, the standard errors of the estimates were very large, making the estimates unreliable. As we noted, the main limitation of short-term studies is that they can make it difficult for the STARTS model to separate the stable trait from the autoregressive components. To avoid this problem, the STARTS model can be modified by collapsing completely stable and relatively stable sources of variance in an effort to distinguish trait-like sources of variance from state variance. This can be accomplished in two ways. One way to modify the STARTS model is to drop the stable trait component from the model, leaving only the autoregressive trait and state components. We refer to this model as the ART-S model. Alternatively, the autoregressive trait component may be dropped from the model, leaving only the stable trait and state components. We refer to this as the ST-S model. As we explain below, the ART-S model should be more useful than the ST-S model for determining the dependability of constructs.

As in the original STARTS model (Kenny & Zautra, 1995), ART-S and ST-S models make a set of assumptions. One assumption is that all of the variance in the latent variable is explained by the two components in the model (i.e., ART and S, or ST and S). Another assumption is that each component explains the same amount of variance at each wave. A third

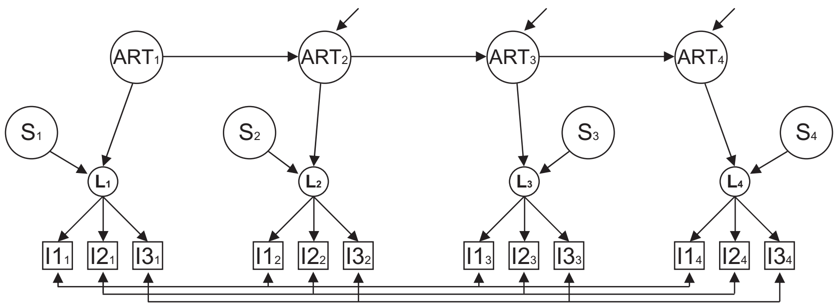

Figure 1
The ART-S model. Ij_n = observed item *j* at Wave *n*; L_n = latent variable at Wave *n*; ART_n = autoregressive trait at Wave *n*; S_n = state component at Wave *n*.

assumption is that the different sources of variance are independent within each measurement occasion and across time. The ART-S model makes an additional assumption that the stability of the autoregressive component over equal intervals does not change. We next describe each of the two models, their advantages, and their limitations in more detail.

ART-S model. The ART-S model is pictured in Figure 1. Two sources of variance influence latent variables in each wave. The autoregressive factors, modeled separately at each wave, represent variance in the latent variable that is partially shared over at least two consecutive waves. Thus, the autoregressive trait at one wave influences the autoregressive trait at the next wave. The estimate of these stability coefficients quantifies the degree of rank-order stability of the autoregressive component from one measurement occasion to the next. In our dataset, the interval between the sixth and seventh assessments was longer than the other retest intervals (2 weeks rather than 1 week, due to spring break). To keep the model simple (i.e., with equal retest intervals), we inserted a “phantom” variable between Waves 6 and 7 to serve as the missing assessment wave.[2] By using

2. A phantom variable was created by constraining its loading from the Wave 1 latent variable to zero. Autoregressive paths from the previous wave to the phantom variable wave, and the phantom variable wave to the following wave, were constrained to be equal to other autoregressive paths between two consecutive waves. This makes the model appear simpler because consecutive assessments are spaced at equal retest intervals. The same results can be obtained by omitting the phantom variable and instead setting the autoregressive path between Waves 6 and 7 to the squared value of other autoregressive paths between two consecutive waves. In both cases, we account for the wave of data that was missing due to spring break.

this method, we could constrain the stability of the autoregressive trait component to be equal across waves, and this single stability estimate can be interpreted as a 1-week stability of the autoregressive trait. The second source of variance in latent variables at each wave is the state factor, which influences responses only at one single measurement occasion. Because of the equality constraints imposed on the model (i.e., amount of autoregressive and state variance is equal at each wave, stability of the ART component is equal over equal intervals), the structural part of the ART-S model estimates just three parameters: (1) the amount of variance at each wave that is explained by the autoregressive trait factor, (2) the amount of variance that is unique to each wave (i.e., state variance), and (3) the 1-week stability of the autoregressive trait.

The main advantage of this approach is that the ART-S model will yield an accurate estimate of the state variance because state variance is defined as any variance that is not shared between any two successive waves. This matches well to the notions of dependability advanced by Chmielewski and Watson (2009) because the retest intervals are short enough to rule out the possibility of real change in the constructs. Thus, the state variance in latent variables will reflect transient errors specific to each measurement occasion. The main limitation of the ART-S model is that the stability of the autoregressive component will be overestimated because the amount of the autoregressive variance will be upwardly biased, as it will include all stable variance that is shared across *all* assessments.

ST-S model. The ST-S model, shown in Figure 2, attributes all of the variance in latent variables at each wave to two factors. The first factor is the stable trait factor, which represents variance in responses that is shared among *all* measurement waves. The second factor is the state factor. As in the ART-S model, the state factor reflects influences that are specific to particular measurement occasions. Given the equality constraints, the ST-S model provides two estimates: (1) the amount of stable trait variance at each wave and (2) the amount of state variance at each wave.

The main advantage of the ST-S model is that it is intuitively simple, as all variance is attributed to either stable or variable factors. However, the model is likely to overestimate the amount of both stable trait and state variance if the construct in question actually contains any autoregressive trait variance. The amount of state variance will be overestimated because some of the autoregressive variance that is not shared across all assessment occasions will be grouped with the state variance. Similarly, the amount of trait variance will be overestimated because some of the autoregressive variance (i.e., that which is present across all

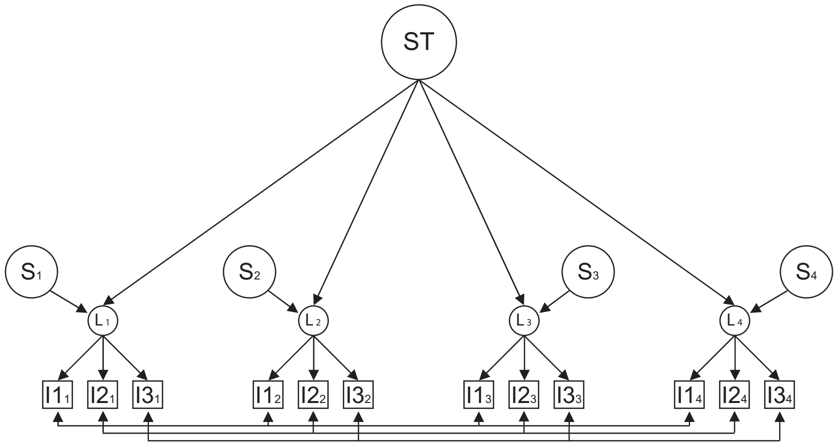

Figure 2
The ST-S model. Ij_n = observed item j at Wave n; L_n = latent variable at Wave n; ST = stable trait; S_n = state component at Wave n.

waves) will be assigned to the trait component. Fortunately, the ST-S model is nested under the ART-S model, as simply constraining the autoregressive path coefficient of the ART-S model to 1.0 would result in the ST-S model. Thus, we can test which of these two models is more plausible using the logic of nested model comparisons in structural equation modeling.

RESULTS

Correlations and descriptive statistics of the scales (based on item means at each wave) are presented in Table 1. The most important point to note from this table is that correlations over equal-length retest intervals (e.g., Week 1 to Week 2, Week 2 to Week 3) remain fairly stable throughout the study. This provides evidence for our assumption that a 1-week interval is long enough to rule out the possibility of memory effects in responding. If participants relied on their memory to answer questions about their personality attributes, we would expect to see higher correlations over equal intervals later on in the study compared to early on. The average 1-week retest correlations were .79 for personality scales (mean r = .78, .83, .76, .70, and .86 for Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness, respectively), .76 for life

Table 1
Correlations and Descriptive Statistics for the Personality, Life Satisfaction, and Affect Scales Over the Eight Waves of Study

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 8	*M*	*SD*
Week 1									
N	—							2.85	0.88
E	—							3.49	0.93
O	—							2.33	0.78
A	—							4.21	0.62
C	—							3.61	0.91
LS	—							5.24	1.10
GD	—							2.02	0.68
AD	—							2.51	0.73
AA	—							1.45	0.48
Week 2									
N	.76	—						2.85	0.87
E	.85	—						3.44	0.90
O	.74	—						2.44	0.74
A	.65	—						4.07	0.66
C	.82	—						3.60	0.90
LS	.72	—						5.42	1.06
GD	.66	—						1.99	0.73
AD	.71	—						2.60	0.72
AA	.63	—						1.33	0.42
Week 3									
N	.76	.76	—					2.77	0.82
E	.85	.87	—					3.48	0.89
O	.73	.73	—					2.37	0.75
A	.67	.67	—					4.07	0.68
C	.82	.84	—					3.67	0.87
LS	.68	.76	—					5.39	1.15
GD	.60	.71	—					1.84	0.73
AD	.68	.72	—					2.52	0.77
AA	.59	.69	—					1.29	0.45
Week 4									
N	.75	.78	.77	—				2.81	0.88
E	.78	.80	.80	—				3.51	0.87
O	.70	.70	.73	—				2.41	0.81
A	.66	.56	.66	—				4.08	0.70
C	.81	.84	.86	—				3.60	0.94
LS	.61	.71	.78	—				5.47	1.13
GD	.60	.60	.68	—				1.81	0.78
AD	.60	.66	.67	—				2.53	0.80
AA	.67	.65	.77	—				1.28	0.50

(Continued)

Table 1 (Cont.)

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 8	*M*	*SD*
Week 5									
N	.69	.71	.81	.79	—			2.77	0.87
E	.78	.81	.85	.78	—			3.45	0.88
O	.73	.72	.75	.74	—			2.43	0.79
A	.65	.58	.75	.73	—			4.04	0.74
C	.82	.79	.84	.86	—			3.67	0.87
LS	.64	.70	.78	.74	—			5.44	1.14
GD	.52	.51	.63	.62	—			1.78	0.74
AD	.56	.60	.67	.61	—			2.53	0.81
AA	.60	.54	.63	.68	—			1.32	0.50
Week 6									
N	.72	.73	.76	.77	.79	—		2.80	0.85
E	.76	.76	.83	.79	.83	—		3.47	0.86
O	.68	.68	.72	.71	.78	—		2.39	0.84
A	.59	.54	.68	.64	.73	—		4.10	0.73
C	.77	.79	.83	.84	.86	—		3.60	0.94
LS	.63	.63	.75	.77	.83	—		5.54	1.11
GD	.55	.58	.66	.60	.68	—		1.80	0.76
AD	.56	.53	.62	.60	.69	—		2.54	0.80
AA	.59	.63	.70	.72	.72	—		1.28	0.46
Week 8									
N	.71	.69	.75	.73	.79	.81	—	2.71	0.84
E	.76	.73	.79	.78	.84	.81	—	3.52	0.88
O	.68	.68	.71	.76	.80	.81	—	2.37	0.82
A	.62	.54	.65	.70	.71	.72	—	4.09	0.74
C	.79	.77	.84	.86	.90	.86	—	3.67	0.91
LS	.52	.50	.65	.57	.67	.72	—	5.61	1.07
GD	.48	.50	.61	.54	.66	.69	—	1.65	0.76
AD	.45	.47	.50	.52	.56	.57	—	2.26	0.78
AA	.49	.51	.50	.60	.71	.66	—	1.22	0.42
Week 9									
N	.73	.78	.78	.79	.83	.83	.84	2.75	0.90
E	.79	.79	.83	.83	.86	.86	.86	3.49	0.88
O	.69	.72	.74	.68	.79	.80	.83	2.33	0.81
A	.62	.58	.67	.73	.75	.72	.77	4.13	0.69
C	.80	.80	.83	.85	.89	.86	.89	3.62	0.91
LS	.58	.60	.68	.65	.75	.78	.74	5.55	1.12
GD	.49	.51	.60	.61	.67	.69	.74	1.75	0.77
AD	.50	.51	.55	.62	.63	.64	.66	2.44	0.79
AA	.54	.53	.58	.69	.75	.76	.76	1.23	0.46

Note. $N = 199–222$. N = Neuroticism; E = Extraversion; O = Openness to Experience; A = Agreeableness; C = Conscientiousness; LS = life satisfaction; GD = general distress; AD = anhedonic depression; AA = anxious arousal.

satisfaction, and .69 for affect scales (mean $r = .68$, .68, and .71 for general distress, anhedonic depression, and anxious arousal, respectively).

Table 2 shows the fit indices and proportions of variance that can be attributed to autoregressive influences and state influences of the ART-S model. In general, the models fit well according to the fit criteria suggested by Hu and Bentler (1999): CFI > .95, RMSEA < .06, SRMR < .08.

Initially, there were two exceptions to the generally good fit: Openness to Experience and Agreeableness. In our original analyses, the fit of the two models for Openness to Experience was below standard level of good fit, ART-S model: $\chi^2(378) = 723.69$, CFI = .923, RMSEA = .064, SRMR = .090; ST-S model: $\chi^2(379) = 741.20$, CFI = .919, RMSEA = .065, SRMR = .091. Similarly, the two models did not fit well for the Agreeableness scale, ART-S model: $\chi^2(378) = 704.07$, CFI = .917, RMSEA = .062, SRMR = .092; ST-S model: $\chi^2(379) = 734.24$, CFI = .910, RMSEA = .064, SRMR = .093.

Further examination revealed that poor fit for these two scales was due to the first step in which the latent factors were specified from scale items. That is, the measurement model, which specified latent variables for Openness or Agreeableness at each wave and simply allowed the latent variables to correlate freely across waves, did not fit the data well. We examined the questionnaire items and found that two items of the Openness scale contained quite similar information and used similar wording (i.e., "Have a vivid imagination" and "Do not have a good imagination"). This led to higher correlations between these items than would be predicted from the basic measurement model, which only allowed items to be related through their association with the common latent trait. Thus, we dropped the partially redundant item "Do not have a good imagination." This substantially improved the fit of the two models, as reported in Table 2. Importantly, the proportion of ART and S variance for Openness did not change with this modification (95% ART, 5% S). Similarly, in the ST-S model, the estimates of stable trait and state proportions did not change (91% ST, 9% S).

We noted a similar issue in the questionnaire items for Agreeableness. Due to their similar wording, the items "Sympathize with others' feelings" and "Feel others' emotions" had higher correlations with each other than with other items. Thus, we decided to drop the

Table 2
Model Fit Indices for the ART-S and ST-S Models, and the Chi-Square Difference Test for the Two Models

	ART-S Model					ST-S Model						
	χ^2	*df*	CFI	RMSEA	SRMR	χ^2	*df*	CFI	RMSEA	SRMR	$\Delta\chi^2$	Δdf
N	558.28	378	.965	.046	.063	567.77	379	.963	.047	.063	9.49	1
E	580.03	378	.968	.049	.049	592.65	379	.967	.050	.050	12.62	1
O[a]	280.73	187	.972	.047	.055	293.88	188	.968	.050	.055	13.15	1
A[b]	303.71	187	.956	.053	.070	311.21	188	.953	.054	.071	7.50	1
C	541.84	378	.975	.044	.044	556.68	379	.973	.046	.044	14.84	1
LS	258.81	187	.987	.041	.059	356.50	188	.969	.063	.075	97.69	1
GD	218.61	187	.994	.027	.063	286.25	188	.980	.048	0.071	67.64	1
AD	245.94	187	.988	.037	.061	311.52	188	.975	.054	.078	65.58	1
AA	293.91	187	.975	.050	.087	348.32	188	.962	.061	.096	54.41	1

Note. All chi-square difference tests are significant at the $\alpha = .05$ level. N = Neuroticism; E = Extraversion; O = Openness to Experience; A = Agreeableness; C = Conscientiousness; LS = life satisfaction; GD = general distress; AD = anhedonic depression; AA = anxious arousal.

[a]The item "Do not have a good imagination" was excluded from these analyses. [b]The item "Sympathize with others' feelings" was excluded from these analyses.

item "Sympathize with others' feelings." This change improved the fit (Table 2) but did not substantially affect the results. The proportions of autoregressive trait and state components remained unchanged after this modification (88% ART, 12% S), and the proportions of the stable trait and state components changed only slightly (82% ST, 18% S in the original model, and 84% ST, 16% S in the modified model).

As the ST-S model is nested within the ART-S model, we were able to compare the chi-square estimates for the two models to examine whether the ART-S model significantly outperformed the ST-S model. As shown in Table 2, this test indicated that the ART-S model fit the data significantly better than the ST-S model for all constructs, reflecting that small systematic changes occur in all three constructs over a period of 2 months. These differences in fit were especially pronounced for life satisfaction and affect. Indeed, one interesting finding to note from Table 2 is that, although the fit of both ART-S and ST-S models was generally adequate across constructs, the discrepancy in fit indices was larger for life satisfaction and affect ratings than for personality ratings. As an example, for Neuroticism, the fit of the ART-S model, CFI = .965, RMSEA = .046, SRMR = .063, was very similar to the fit of the ST-S model, CFI = .963, RMSEA = .047, SRMR = .063. In contrast, for life satisfaction, the difference in fit between the two models was more apparent, ART-S model: CFI = .987, RMSEA = .041, SRMR = .059; ST-S model: CFI = .969, RMSEA = .063, SRMR = .075. This suggests that personality is stable enough over 2 months to be more or less adequately described by the ST-S model. In contrast, life satisfaction and weekly reports of affect are more sensitive to changes that accumulate over time, and thus the ART-S model is more appropriate for their description. In short, Table 2 revealed some important differences between personality, on one hand, and life satisfaction and affect on the other.

The strength of the autoregressive effects on a construct can also be gleaned from the discrepancy in variance allocated to the S component in the two models. In the ART-S model, the S component captures any variance that is not shared between two adjacent waves, whereas in the ST-S model it captures the variance that is not shared among all waves. Because of this, an estimate of S that is substantially larger in the ST-S model than in the ART-S model would be indicative of variance that is shared between some adjacent waves,

Table 3
Proportion of Variance Accounted for by Each of the Components in the ART-S and ST-S Models

	ART-S Model			ST-S Model	
	Stability	ART	S	T	S
N	.990	91%	9%	89%	11%
E	.991	92%	8%	90%	10%
O[a]	.981	95%	5%	91%	9%
A[b]	.983	88%	12%	84%	16%
C	.992	96%	4%	94%	6%
LS	.952	86%	14%	75%	25%
GD	.952	76%	24%	65%	35%
AD	.951	75%	25%	65%	35%
AA	.957	80%	20%	69%	31%

Note. N = Neuroticism; E = Extraversion; O = Openness to Experience; A = Agreeableness; C = Conscientiousness; LS = life satisfaction; GD = general distress; AD = anhedonic depression; AA = anxious arousal; stability = stability of the autoregressive trait; ART = autoregressive trait; T = trait; S = state.
[a]The item "Do not have a good imagination" was excluded from these analyses.
[b]The item "Sympathize with others' feelings" was excluded from these analyses.

but not all waves. This would suggest that autoregressive trait is an important component for that construct. Variance component estimates, shown in Table 3, offer convergent evidence that short-term autoregressive influences are stronger for life satisfaction and affect than personality over a 2-month period. Discrepancy in estimates of S between the two models is small for personality (average difference = 3%) and larger for life satisfaction (difference = 11%) and affect (average difference = 11%). Because the S estimates for personality are roughly equal in the two models, we can conclude that the autoregressive influences likely have only minor effects on personality ratings over 2 months. On the other hand, the S estimates are substantially larger for the ST-S than ART-S model for both life satisfaction and affect, suggesting that autoregressive effects are stronger for these two constructs. These results suggest that the changes in life satisfaction and affect observed over a period of 2 months are due in part to changes in life circumstances during this period. In contrast, changes in personality ratings over 2 months can be mostly attributed to transient influences specific to a single measurement occasion. The important point is that such insights are only gained when both the ST-S and ART-S model are fit to the same data.

Similar conclusions can be drawn from the estimates of the stability of the autoregressive component. This stability is lower for life satisfaction and affect than for personality. Although this difference may seem small, it is important to keep in mind that these are 1-week stabilities, and that small changes can accumulate over time into meaningful differences in the extent to which we can predict different constructs over time, such as a 2-month period. For example, the results of the ART-S model predict that 83% of the autoregressive trait variance in Neuroticism is still present after the 9 weeks of our study ($.990^9 = .91$, $.91^2 = .83$). On the other hand, only 41% of autoregressive trait variance in life satisfaction is still present after this time period ($.952^9 = .64$, $.64^2 = .41$). This finding can be interpreted in two ways. First, it is possible that the autoregressive component is simply more stable for personality than for life satisfaction and affect. Alternatively, this finding can also be interpreted as evidence that personality is influenced more by stable trait factors than by autoregressive factors, because the stable trait component present in personality ratings would inflate stability estimates for the autoregressive component. This interpretation is also consistent with the evidence that the autoregressive component is more important for life satisfaction and affect than for personality.

Finally, it can be noted in Table 3 that for both ART-S and ST-S models, proportion of state variance is lowest for personality and highest for affect. Because these models use latent variables for each wave, the state component reflects reliable occasion-specific variance that is not associated with variance at any other wave. This finding suggests that personality measures are the least affected by these transitory sources of variance, whereas measures of affect are most affected. This also makes sense given that life satisfaction measures and especially the affective symptom reports are thought to be at least somewhat responsive to quickly changing life circumstances.

DISCUSSION

This article illustrated how the STARTS model can be applied to short-term longitudinal data to provide important conceptual and methodological insights into the study of the consistency of psychological attributes over time. Specifically, we compared the stability and change in personality, life satisfaction, and affect over a period of 2 months. Our goal was to isolate the stable and changing elements of three classes of constructs that should vary in their stability over time. This information is conceptually important in its own right, as it suggests that even short-term studies can help position constructs on the continuum of consistency. Methodologically, this information can inform other studies by providing more precise evidence regarding the dependability of the measures (Chmielewski & Watson, 2009).

In this work, we showed how Kenny and Zautra's (2001) STARTS model can be applied to short-term longitudinal data to thoroughly evaluate the dependability of constructs. In particular, we tested two models: one that assumed that variance in individual differences is either fully stable or specific to each measurement occasion (ST-S model) and another that allowed for accumulation of changes from one assessment to the next in addition to occasion-specific influences (ART-S model). When comparing these two models across the three sets of constructs we examined, clear differences in fit emerged. Indeed, we generally recommend that researchers fit both models and compare results. In these data, the ART-S models fit better than the ST-S model; however, we believe that the ST-S model may still be useful for modeling personality data over short time periods. One reason is that the ST-S model is simpler, more intuitive, and generally easier to model than the ART-S model. In addition, estimates of model fit and transient influences on personality did not differ substantially for the two models, and thus researchers would draw similar conclusions regardless of the model they employed in their studies. Thus, a model that includes only a stable trait component and transitory influences provides a relatively parsimonious description of personality trait measures over a 2-month period.

This was not the case, however, for measures of life satisfaction and affect. For these constructs, the ST-S model fit much worse than the ART-S model, and the interpretations of the models would differ substantially across the two measures. Specifically, in both cases, the transitory state component is exaggerated when the ST-S model is fit to the data. The ART-S model showed that some of the variance that is partitioned into the state component in the ST-S model is actually autoregressive variance that is not completely stable over the course of the 2-month period, but it does change relatively slowly over this interval. Again, this suggests that transitory influences would be overestimated if the autoregressive trait influences were not taken into account. Thus, life satisfaction and affect are best described using the ART-S model, whereas the ST-S model is likely to be adequate for describing personality traits in short-term longitudinal studies.

Another important finding was that the amount of transitory influences differed across constructs. Our results indicated that, on average, 8% of reliable variance in personality measures was due to transient influences. In contrast, such influences accounted for 14% of variance in life satisfaction judgments. Thus, state factors seem to influence life satisfaction more than they affect personality, which suggests that transient states such as mood or events specific to a single week are reflected in life satisfaction ratings more so than in ratings of personality. On the other hand, ratings of affect are even more influenced by state factors (23%). This is not surprising when one considers that in this study, judgments of affect were made in relation to the past week only, and not in general. Thus, influences specific to a single week are expected to play a larger role in these judgments than in general ratings of personality or life satisfaction.

In future studies it may be important to separate stable from transient variance in individual differences, depending on research questions of interest. For example, if a researcher is interested in causes and effects of transient influences on questionnaire ratings (e.g., temporary response styles or mood effects), she or he may wish to isolate the state component and examine its relations with other variables. More often, personality psychologists are interested in nontransient variance in individual differences, in which case it is important to remove state influences at each wave and examine causes and effects of the dependable portion of variance.

Our results have important implications for analyses of the dependability of measures. Namely, the results suggest that different retest intervals may be appropriate for different constructs. Personality measures seem to be largely trait-like over a 2-month period, whereas life satisfaction and affect show true change over 2 months. Thus, a 2-month retest interval may be suitable for dependability analysis of personality measures, whereas a shorter retest interval (e.g., 1 week) seems to be more appropriate for life satisfaction and affect. Using longer retest intervals for dependability analysis of life satisfaction and affect may lead to overestimates of transient error, as retest correlations would be attenuated not only by transient influences but also true autoregressive changes in these two constructs.

In sum, using the modified STARTS models to establish appropriate retest intervals for dependability analyses has a number of advantages over using simple short-term retest correlations. First, by fitting our models to latent variables, we were able to separate random measurement error from other sources of nondependability. This is particularly important for constructs that have different reliabilities. If random measurement error and other transitory influences are confounded, it is not possible to judge whether differences in dependability estimates between constructs are due to differences in reliabilities, differences in transient influences, or both. Second, by comparing the ART-S and ST-S models, we were able to determine whether a construct showed true autoregressive change over the period of our study. This led us to make our conclusions about the appropriateness of different retest intervals for dependability analyses. Third, using our modified models in future research, it is possible to separate different sources of variance and examine their causes and effects in their own light. For example, it is possible to further examine the nature of transitory influences on individual differences by examining correlates of the state component from the ART-S model.

Limitations and Future Directions

Because of the short-term nature of our dataset, we were not able to examine the full STARTS model. Undoubtedly, the full STARTS model would give us the most accurate estimates of the different sources of variance in individual differences over time. However, because of the relatively stable nature of individual differences in our study, the full STARTS model would likely require data assessments over a much longer period of time so that the autoregressive components could approach zero when multiplied over enough intervals. Nonetheless, we believe that the information provided by the ART-S and ST-S models allowed us to answer some important questions about dependability and stability that were practically useful.

Using the weekly retest intervals in our study, we were able to separate transitory influences from other more long-lasting changes in our constructs. One limitation of this method is that memory effects can distort our results, as participants may make an effort to answer in a consistent manner across the waves. Other studies can test this hypothesis by manipulating retest intervals or using a planned missing data design whereby participants in the same longitudinal study are selected at random to give weekly responses. However, it is important to note that differences in transitory influences were observed across the three constructs, and it is unlikely that differences in memory for the three sets of constructs could explain these effects, as we would expect that memory effects would be constant across all items. In addition, examination of correlations across waves suggests that memory effects were minimal since responses did not become more consistent over the course of the study.

Another limitation of our study is that we were able to include only one measure of each construct. Thus, dependability findings may be measure-specific, rather than informative about transient effects on broad constructs of personality, life satisfaction, and affect. We were, however, able to separate random measurement errors, which are likely to be scale-specific, from systematic transient influences, which may generalize across scales. Future studies can compare the effects of transient influences across different scales in order to distinguish between transient factors that affect specific measures and those that apply to broad constructs. Similarly, future studies can test whether our findings about the extent of transient influences in the three constructs generalize to nonstudent populations. For example, life circumstances of a college student may be more volatile compared to those of a middle-aged adult whose family and career situations are undoubtedly more stable. Thus, the dependability estimates may change over the course of development.

In closing, the study of consistency of individual differences is likely to continue to hold a central place in personality psychology for the foreseeable future. What will likely change are the methods and models used to study this topic, and we expect to see increasing methodological (e.g., Chmielewski & Watson, 2009; Kenny & Zautra, 2001; Watson, 2004) and conceptual sophistication (e.g., Fraley & Roberts, 2005) in this area. In line with these trends, the current study demonstrated two ways in which the STARTS model may be modified for applications to short-term longitudinal data (i.e., data that span months rather than years). We also showed that for certain constructs (i.e., life satisfaction, affect), it may be important to model the influences of the autoregressive trait in order to capture all of the valid (i.e., not occasion-specific) variance. We predict that STARTS models and other trait-state models (e.g., Cole, 2006; Steyer et al., 1999) will become particularly useful and important tools for conceptualizing and empirically investigating stability and change in psychological research.

REFERENCES

Anusic, I., & Schimmack, U. (2011). *Evidence for stable personality factors: A meta-analytic model of retest correlations*. Unpublished manuscript.

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, **97**, 186–202.

Clark, L. A., & Watson, D. (1995). *The Mini Mood and Anxiety Symptom Questionnaire (Mini-MASQ)*. Unpublished manuscript.

Cole, D. A. (2006). Coping with longitudinal data in research on developmental psychopathology. *International Journal of Behavioral Development*, **30**, 20–25.

Conley, J. J. (1984). The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality and self-opinion. *Personality and Individual Differences*, **5**, 11–25.

Cooper, A. J., Smillie, L. D., & Corr, P. J. (2010). A confirmatory factor analysis of the Mini-IPIP five-factor model personality scale. *Personality and Individual Differences*, **48**, 688–691.

Costa, P. T., Jr., & McCrae, R. R. (1986). Personality stability and its implications for clinical psychology. *Clinical Psychology Review*, **6**, 407–423.

Costa, P. T., Jr., & McCrae, R. R. (1997). Longitudinal stability of adult personality. In R. Hogan, J. A. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 269–290). San Diego, CA: Academic Press.

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

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**, 192–203.

Ferguson, C. J. (2010). A meta-analysis of normal and disordered personality across the life span. *Journal of Personality and Social Psychology*, **98**, 659–667.

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**, 60–74.

Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review*, **34**, 93–101.

Hertzog, C. & Nesselroade, J. R. (1987). Beyond autoregressive models: Some implications of the trait-state distinction for the structural modeling of developmental change. *Child Development*, **58**, 93–109.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, **6**, 1–55.

James, W. (1890). *Principles of psychology*. New York: Henry Holt.

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology*, **63**, 52–59.

Kenny, D. A., & Zautra, A. (2001). The trait-state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 243–263). Washington, DC: American Psychological Association.

Low, K. S. D., Yoon, M., Roberts, B. W., & Rounds, J. (2005). The stability of vocational interests from early adolescence to middle adulthood: A quantitative review of longitudinal studies. *Psychological Bulletin*, **131**, 713–737.

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

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 92–105). Washington, DC: American Psychological Association.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, **126**, 3–25.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, **13**, 389–408.

Terracciano, A., McCrae, R. R., Brant, L. J., & Costa, P. T., Jr. (2005). Hierarchical linear modeling analyses of the NEO-PI-R scales in the Baltimore Longitudinal Study of Aging. *Psychology and Aging*, **20**, 493–506.

Trzesniewski, K. H., Donnellan, M. B., & Robins, R. W. (2003). Stability of self-esteem across the life span. *Journal of Personality and Social Psychology*, **84**, 205–220.

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, **38**, 319–350.