Fall 2013

Transient Measurement Error in a Diverse Population

Sean Potter
Southern Methodist University, spotter@smu.edu

Transient Measurement Error in a Diverse Population

Sean Potter, Dr. Michael Chmielewski

Southern Methodist University

Transient Measurement Error in a Diverse Population

The accurate assessment of the stability of personality traits is important for psychological researchers and society in general. In the clinical world, personality trait stability has importance for diagnosing clinical disorders, especially personality disorders such as borderline personality disorder. Trait stability also influences the effectiveness and necessity of therapeutic interventions (Costa & McCrae, 1997). Moreover, evidence regarding trait stability can provide insight into the very nature of personality itself (Costa & McCrae, 1997). Accurately assessing personality stability is also important for our workplace because personality measures are used to study a variety of functions in the context of our workforce. However, it is especially concerning to learn that measurement error is a widespread phenomenon that affects all areas of science (Schmidt, Le, & Ilies, 2003) and that personality researchers have not given measurement error the attention it deserves (Chmielewski & Watson, 2009; Watson, 2004).

Currently, the golden standard to determine the impact of measurement error in personality assessment is the Cronbach's Alpha formula (Cronbach, 1951). However, Cronbach's Alpha does not take into account the influence of transient measurement error influencing a measure's reliability (Chmielewski & Watson, 2009; Schmidt et al., 2003; Watson, 2004). Transient errors are caused by fluctuations in participants' psychological states at a particular assessment (Chmielewski & Watson, 2009; Schmidt et al., 2003; Watson, 2004). These state fluctuations can then have a substantial impact on how individuals' respond to trait measures (Chmielewski & Watson, 2009). How much variance in personality measures can be explained by the role of transient error? Chmielewski and Watson (2009) found that nearly 25% of the variance in trait measures was due to transient error. This has led to an increased interest in transient measurement error (Anusic, Lucas, & Donnellan, 2012). However, studies investigating transient error have relied solely on data from college students who have typically been Caucasian and 18 years old. In addition, prior studies have also found a pattern that suggests the wording or item formatting of measures may influence transient error. Chmielewski and Watson (2009) found in their study the BFI, a measure for the Five Factor Model of personality that uses sentence formatting for items item, had generally lower levels of transient error compared to its counterpart the Goldberg Five Factor markers, a measure that uses single word formatting for items. This same pattern emerged in the same study with the TEQ, an affect measure that uses sentence formatting, and the PANAS-X, an affect measure that uses single word formatting.

The purpose of this project is to help further analyze this phenomenon of transient error. Specifically, I will examine whether transient error influences personality assessment in an older, more diverse sample. We do know that personality traits, although fairly stable throughout most of our lives, grow even more stable as individuals age (Costa & McCrae, 1997). As such, it possible that transient error may be a serious issue only when assessing younger adults as their views of their personality may not be as stable. In addition, the majority of participants in past studies have been Caucasian and it is unknown if the results can be generalized to other groups. Our hypothesis is that compared to prior samples, levels of transient error will lower in this older, more diverse sample but still at a level to cause concern for researcher.

Procedure

Participants (n=480) were users registered with Amazon's Mechanical Turk (MTurk). It is an online service where people sign up to complete various tasks, including tasks for research studies, in exchange for compensation. Data that is collected via MTurk has proven to be similar to data collected via more traditional methods (Buhrmester, Kwang, & Gosling, 2011). Participants completed a one-week test-retest of the same measures. A short test-retest interval of one week was used to control for the possibility of significant life events occurring (death in the family, major illness, etc.) that could produce actual personality change. Within one week, any differences in responses can reasonably be attributed to transient measurement error and not true trait change. Attention checks (a total of 16 for time 1 and 15 for time 2) were included with every measure that asked participants to select a certain response. For example, one attention check asked "Please select strongly disagree". The measures used in the study are listed below and roughly grouped around the research domains they are frequently used. However, these measures are used across the disciplines commonly and not restrained in their use to any domain.

Demographics

A key goal of this study was to collect data on participants who are more diverse than the typical undergraduate samples collected for personality studies. The average age of the participants was 38.5 with ages ranging from 18 to 75. Males constituted 28.5% of the sample and women made up 63.5% of the sample. Most participants reported being White (76%), followed by Black (6.7%) and Asian (6.3%). Most participants reported having some college education (24%), a full college education (31%), some graduate education (3.3%) or a full graduate degree (16%). The rest reported having either vocational education (1.5%), graduating high school (11.0%) or haven't completed high school (1.3%). Most participants reported making less than $25,000 (34.8%), followed by between $25k-$50k (29.4%), $50k-$75k (17.7%), $75k-$100k (4.4%) and more than $100k (0.2%). Overall, this sample has a greater age, educational, socioeconomic, and racial diversity than what generally undergraduate samples have achieved in prior samples (Chmielewski & Watson, 2009).

Personality Psychology Measures

The PANAS-X (Watson & Clark, 1994) is a factor analytically derived 60-item measure of affectivity. The trait version of the instrument was used, which asks participants to indicate on a 5-point scale ranging from 1 (very slightly or not at all) to 5 (extremely) "to what extent you generally feel this way, that is, how you feel on the average." Scales included for analyses were two higher order scales, General Negative Affect and General Positive Affect, and 9 of the PANAS-X specific affect scales (Shyness, Fatigue, Serenity, Hostility, Guilt, Sadness, Joy, Self-Assurance) were included that measure specific types of affect.

The Temperament and Emotion Questionnaire (TEQ; Watson, 2004) is a 60-item measure of affectivity created by embedding the PANAS-X descriptors into complete sentences. For example, the PANAS-X item "cheerful" became "I am a cheerful person." Participants rate each statement on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree) .The TEQ contains the same scales as the PANAS-X. The TEQ Shyness, Fatigue and Serenity scales were newly created for the 2-week retest study; all of the other TEQ scales are also available in the 2-month sample. The convergent correlations between the parallel PANAS-X and the TEQ scales ranged from .57 to .80 across the various time points (grand M = .71).

The Big Five Inventory (BFI; John & Srivastava, 1999) is a widely used, factor analytically derived, 44-item measure of the five-factor model of personality, assessing Neuroticism, Extraversion, Agreeableness, Conscientiousness, and Openness. The BFI includes eight-item Neuroticism and Extraversion scales, nine-item measures of Agreeableness and

Conscientiousness, and a 10-item Openness scale. The instructions include an initial statement that reads “I see myself as someone who… ”; participants then read each item (e.g., “is talkative“) and responded on a 5-point scale ranging from 1 (disagree strongly) to 5 (strongly agree). The BFI is available in all samples.

To create a second set of Big Five scales, 45 adjectives were selected from Goldberg’s (1992) list of Big Five factor markers. Nine items each were chosen for Neuroticism, Extraversion, Agreeableness, Conscientiousness, and Openness to coincide with the BFI scales. Participants rated themselves on each adjective on a 5-point scale ranging from 1 (very inaccurate) to 5 (very accurate) with regard to how well the term described them.

Social Psychology Measures

The 10-item Rosenberg self-esteem scale (RSES; Rosenberg, 1965) was used to measure participants’ level of self-esteem. Participants responded on a 4-item likert scale ranging from strongly disagree to strongly agree.

The Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985) is a 15-item instrument that measure judgments of satisfaction with one’s life. Responses are placed on a 7-point likert scale ranging from “Strongly Disagree” to “Strongly Agree”. The SWLS has three subscales: satisfaction with one’s past life, present life, and future life.

The Subjective Happiness Scale (SHS; Lyubomirsky, & Lepper, 1999) is a 4-item measure that assesses global subjective happiness. Participants answered each question on 7-point scales.

Workplace Psychological Measures

We used Rotter's 29-item scale to assess participant locus of control. The scale is forced-choice format where participants must pick which of two statements they agree with most for item. An example pair participants picked between is "Many of the unhappy things in people's lives are partly due to bad luck" and "People's misfortunes result from the mistakes they make".

The Core Self-Evaluation Scale (CSES; Judge, Bono, & Thoresen, 2003) is a 20-item measure that assesses an individual's evaluations about themselves, their own abilities, and their own control. Responses were on a 5-item Likert scale ranging from "Strongly Disagree" to "Strongly Agree".

Finally, the Narcissistic Personality Inventory-16 (NPI-16; Ames, Rose, Anderson, & Cameron, 2006) is a 16-item measure for subclinical narcissism. Its items are drawn from the NPI-40, a 40-item measure of Narcissistic personality (Raskin & Terry, 1988). For each item, participants had to select between a pair of statements that most closely described their feelings and beliefs.

Cronbach's Alpha Reliability Results

In table 1 below, Cronbach's Alpha for each measure at both time 1 and time 2 are listed below. Cronbach's Alpha, since its initial inception, has become the primary standard for determining a measure's reliability (Cronbach, 1951). A popular rule of thumb is Cronbach's Alpha coefficients of 0.70 and above mean the measure has good reliability (Nunnally, 1978). As listed below, all of the measures and measure subscales had coefficients ranging between 0.74 for the TEQ Attentiveness subscale to 0.96 for the PANAS-X Joy subscale. By conventional standards, these measures would be considered to have good reliability by most researchers. However, the Cronbach's Alpha formula does not take into the influence of transient measurement error influencing a measure's reliability (Chmielewski & Watson, 2009; Schmidt et al., 2003; Watson, 2004). In order to determine the impact on reliability from transient error, each measure's dependability was also calculated.

Table 1: Scale Alphas for Time 1 and Time 2

Scale	**Alpha Time 1**	**Alpha Time 2**	**Scale**	**Alpha Time 1**	**Alpha Time 2**
Big Five Extraversion	0.90	0.90	Goldberg Surgency	0.90	0.91
Big Five Agreeableness	0.85	0.86	Goldberg Agreeableness	0.88	0.88
Big Five Conscientiousness	0.88	0.89	Goldberg Conscientiousness	0.90	0.90
Big Five Neuroticism	0.92	0.92	Goldberg Emotional Stability	0.86	0.87
Big Five Openness	0.86	0.87	Goldberg Intellect	0.79	0.81
PANAS-X Positive Affect	0.92	0.92	TEQ Positive Affect	0.85	0.87
PANAS-X Negative affect	0.93	0.94	TEQ Negative affect	0.91	0.82
PANAS-X Attentiveness	0.84	0.87	TEQ Attentiveness	0.74	0.80
PANAS-X Shyness	0.90	0.90	TEQ Shyness	0.90	0.91
PANAS-X Fatigue	0.92	0.94	TEQ Fatigue	0.90	0.92
PANAS-X Serenity	0.88	0.90	TEQ Serenity	0.78	0.82
PANAS-X Fear	0.92	0.93	TEQ Fear	0.91	0.92
PANAS-X Hostility	0.89	0.90	TEQ Hostility	0.85	0.86
PANAS-X Guilt	0.94	0.94	TEQ Guilt	0.90	0.92
PANAS-X Sadness	0.93	0.93	TEQ Sadness	0.92	0.93

PANAS-X Joy	0.95	0.96	TEQ Joy	0.92	0.93
PANAS-X Self-Assurance	0.87	0.88	TEQ Self-Assurance	0.80	0.81
Satisfaction with Life: Past	0.89	0.90	Rosenberg Self-Esteem	0.94	0.93
Satisfaction with Life: Present	0.93	0.94	Locus of Control	0.84	0.86
Satisfaction with Life: Future	0.92	0.93	Core Self-Evaluations	0.84	0.84
Subjective Happiness Scale	0.92	0.92	Narcissistic Personality Inventory-16	0.80	0.81

Dependability Results

In table 2 below, the dependability for each measure and measure subscale is listed. Dependability is calculated by correlating responses for each measure from time 1 and time 2. Because there was only a 1-week retest week interval, any amount of true personality change should be negligible and have no affect on responses for each measure. The short retest interval minimized the possibility of significant life events also affecting participants' responses. Correlations less than 1.0 (with a correlation of 1.0 meaning 100% consistency) indicate that there are inconsistencies in responses to the same measure from time 1 and time 2, with smaller correlation values indicating greater inconsistencies. Because no true change should have occurred to cause changing responses to these trait and trait-like measure, correlations less than 1.0 can be attributed to the influence of transient error and provide a metric of transient error's impact on reliability. For example, the TEQ Attentiveness subscale has a dependability value of 0.749. This value can be interpreted as about 25% (calculated by subtracting 0.749 from 1) of the variance in the TEQ Attentiveness subscale can be attributed to transient measurement error.

Table 2: Scale Dependabilities

Scale	Dependability	Scale	Dependability
BFI Extraversion	0.918	Goldberg Surgency	0.945
Big Five Agreeableness	0.906	Goldberg Agreeableness	0.871
Big Five Conscientiousness	0.912	Goldberg Conscientiousness	0.911
Big Five Neuroticism	0.918	Goldberg Emotional Stability	0.907
Big Five Openness	0.894	Goldberg Intellect	0.855
PANAS-X Positive Affect	0.856	TEQ Positive Affect	0.866
PANAS-X Negative affect	0.882	TEQ Negative affect	0.992
PANAS-X Attentiveness	0.719	TEQ Attentiveness	0.749
PANAS-X Shyness	0.844	TEQ Shyness	0.871
PANAS-X Fatigue	0.837	TEQ Fatigue	0.854
PANAS-X Serenity	0.812	TEQ Serenity	0.829
PANAS-X Fear	0.839	TEQ Fear	0.908
PANAS-X Hostility	0.828	TEQ Hostility	0.853
PANAS-X Guilt	0.878	TEQ Guilt	0.898
PANAS-X Sadness	0.83	TEQ Sadness	0.9
PANAS-X Joy	0.886	TEQ Joy	0.89
PANAS-X Self-Assurance	0.892	TEQ Self-Assurance	0.853
Satisfaction with Life: Past	0.818	Rosenberg Self-Esteem	0.844
Satisfaction with Life: Present	0.865	Locus of Control	0.886
Satisfaction with Life: Future	0.787	Core Self-Evaluations	0.892
Subjective Happiness	0.906	Narcissistic Personality	0.898

Scale	Inventory-16

Overall, the average dependability for the measures was higher than anticipated. Chmielewski and Watson (2009) found in their study on average, 25% of the variance in assessed trait measures was attributed to transient error. In contrast, about 13% of the variance in trait and trait-like measures we collected data on is attributed to transient error. It is interesting to note there the BFI had a higher overall dependability scores for among its five scales compared to its equivalent Goldberg's Five Factor scales. A similar pattern emerged with the TEQ having higher overall dependability scores compared to the PANAS-X. The BFI and Goldberg Five Factor measure both assess essentially the same traits (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness). The TEQ and PANAS-X also measure the same trait-like affective constructs. The difference between these measures is that the BFI and TEQ have each item embedded as a sentence while the Goldberg and PANAS-X both use single word items. Chmielewski and Watson (2009) also found a similar pattern where the BFI and TEQ had higher dependabilities than the Goldberg and PANAS-X respectively.

Discussion and Direction for Future Research

Overall, it appears that transient error may not have as great of an effect on more diverse populations beyond undergraduate populations than was expected. This is good news especially for clinicians and the workplace, both areas that work primarily with older, more diverse populations. Transient error may in fact have less influence on their measures and their results less influenced by measurement error. The reason for this is unclear. One likely possibility is that because our population was older on average than most undergraduate samples, age may play a role in influencing transient error. Our results could have occurred due to the unique attributes of the Mechanical Turk participant population and different results may occur if the same study was administered to an equally diverse community sample. It is interesting to note that similar findings occurred in this study and Watson and Chmielewski's (2009) in that the BFI and TEQ performed better than their single-word item (Goldberg and PANAS-X respectively) in both studies. It may be that the formatting of measures may play a significant role in influencing the impact of transient error.

Future research should explore possibilities for differences in transient error levels between this study and prior studies. Collecting data from Mechanical Turk while focusing primarily on recruiting younger, college-aged participants would be an excellent first step to see that sample has similar levels of transient error to what other undergraduate samples have seen. Future studies should also explore ways of reducing levels of transient error in undergraduate samples.

References

Ames, Daniel R., Rose, Paul, and Anderson, Cameron P. (2006). The NPI-16 as a short measure of narcissism. *Journal of Research in Personality, 40*, 440-450.

Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes Across Cultures and Ethnic Groups: Multitrait Multimethod Analyses of the Big Five in Spanish and English. *Journal Of Personality & Social Psychology*,*75*(3), 729-750.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?. Perspectives On Psychological Science (Sage Publications Inc.), 6(1), 3-5.

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

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

Cronbach, J. (1951). "Coefficient alpha and the internal structure of tests". Psychometrika 16 (3): 297–334.

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

John, O. P., Donahue, E. M., & Kentle, R. L. (1991*).* *The Big Five Inventory--Versions 4a and 54.* Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., Naumann, L. P., & Soto, C. J. (2008*). Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues.* In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), Handbook of personality: Theory and research (pp. 114-158). New York, NY: Guilford Press.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2003). The Core Self-Evaluations Scale: Development of a measure. *Personnel Psychology,* 56(2), 303-331.

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

Nunnally J. C. (1978*). Psychometric Theory,* 2nd ed. New York: McGraw-Hill.

Pavot, W., Diener, E., & Suh, E. (1998). The Temporal Satisfaction With Life Scale. *Journal Of Personality Assessment*, *70*(2), 340-354. doi:10.1207/s15327752jpa7002_11

Saucier, G. (1994). Mini-Markers: A Brief Version of Goldberg's Unipolar Big-Five Markers. *Journal Of Personality Assessment*, *63*(3), 506.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, 8, 206–224.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General And Applied*, *80*(1), 1-28. doi:10.1037/h0092976

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.

Watson, D., & Clark, L. A. (1994*). The PANAS-X: Manual for the positive and negative affect schedule-Expanded Form*. Iowa City: University of Iowa