Test-Retest Reliability of Self-Reported Lifetime History of Aggression: A Unique Longitudinal Assessment Over One Year Using Mechanical Turk

Jenna Kilgore

Department of Psychology, Mississippi State University

Introduction

Defining Reliability

The concept of reliability stems from attempting to measure human qualities with consistency (Kaplan & Saccuzzo, 2013). Classical test theory, which is responsible for the traditional conceptualization of reliability in the discipline of psychology, relates the discrepancy between the observed score (the measured score of a psychological assessment) and the true score, which is the hypothetical actual characteristic of a human being. Variations between these scores are considered to be random – that is, it is assumed that perfect adherence between observed and true scores is possible – so it is the researcher's responsibility to ensure that the measurement error of a psychological assessment is not unduly large.

Reliability is evaluated in different ways within the field of psychology. For example, it can be measured using methods of parallel forms, internal consistency, or test-retest reliability. Parallel, or equivalent, forms of reliability are when different forms of the specific measure are evaluated to ensure they are assessing the same characteristics. That is, researchers must make sure that a measure is representative of an entire characteristic or trait and not only a portion or subgroup of such. This particular type of reliability is not as common as the others, mainly because it is time consuming to develop two separate measures and assess reliability for both (Kaplan & Saccuzzo, 2013).

Internal consistency evaluates items within a measure regarding whether they assess for the same characteristic or ability. There are several ways to measure internal consistency, including the split-half method, Kuder-Richardson method, and coefficient alpha (Kuder & Richardson, 1937, as cited by Kaplan & Saccuzzo, 2013; Cronbach, 1951,

at, the infliction of injury or discomfort; also manifestations of inner reactions such as feelings or thoughts that can be considered to have such an aim are regarded as aggressive responses." Thus, aggression can be seen as an internal or external act, thought, or feeling that is aimed to inflict harm.

Relating aggression to a larger social context, Crick and Dodge (1994) developed the Social Information Processing (SIP) theory that includes cognitive steps used to process social cues that could potentially result in an aggressive response. These steps include (1) encoding cues; (2) interpreting cues; (3) clarifying goals; (4) accessing potential responses; (5) selecting a response; and (6) enacting the response.

Some psychologists, such as Nestor (2002), would place more emphasis on personality dimensions, stipulating that aggression and violence are a result of a lack of impulse control and affect regulation. Further considering aggression in relation to personality, aggression can be conceptualized and measured with either trait- or state-level items (Coccaro et al, 1991; 1997). Trait characteristics are typically defined as characteristics that persist over the lifetime, are fairly stable, and can be seen across a broad spectrum of situations (Winer, Veilleux, & Ginger, 2014). State characteristics, however, are not stable over the lifetime and can change with time and situation. Coccaro et al. (1997) argue that state measures of aggression typically cannot be coupled with aspects of biological analysis. Moreover, state-based measures of aggression would vary highly when examined longitudinally, leading to results that show the same participant as aggressive at one time point and not aggressive the next hour, day, or week. This may be useful in situations where the variability of aggressive behaviors is of

in-person interviews and a review of any clinical data on the participants. The study provided sufficient evidence of adequate psychometric properties of the scale.

Test-Retest Reliability of the LHA

Coccaro and colleagues' original LHA paper has been cited 21 times, as of April 7th, 2016, when searched via the PsycInfo Database. Thus, it is established as a measure of aggression cited with mild regularity in the large aggressive behavior literature. Despite it having been cited a number of times, however, there has been limited work examining the test-retest reliability of the measure.

Indeed, in the original psychometric validation paper (Coccaro et al., 1997), the only test-retest reliability examination of the LHA assessed 20 subjects via two separate in-person interviews ranging from 28 to 360 days apart. Spearman correlations were conducted to assess test-retest reliability over those periods, resulting in test-retest reliabilities for LHA total (r=.91), other-directed aggression (r=.80), consequences/antisocial behavior (r=.89), and self-directed aggression (r=.97) (Coccaro et al, 1997).

Multiple studies have used the 11-item Life History of Aggression scale via semi-structured interview or self-report as the single aggression measure, while others have paired it with other measures of aggression, such as the Buss-Perry Aggression Questionnaire (Berman et al., 1998; Fanning et al., 2014; McCloskey and Berman, 2003; Schoenleber, et al., 2011; Swogger, et al., 2014; Tcheremissine, et al., 2005; Verona, et al, 2012; 2012; Victor and Klonsky, 2014). Other studies did not use the full LHA scale, but instead utilized the individual subscales or shorter versions (9-item or 10-item), such as the other-directed aggression subscale (Berman et al., 2009; Bresin et al., 2013; Fanning

Participants were re-contacted via their de-identified MTurk ID first using Python, then using R software, to ensure complete confidentiality (Mueller & Chandler, 2012; Leeper, 2014). Python was used to re-contact participants only for the second wave. As outlined by Mueller & Chandler (2012), to use Python researchers must download and install Python, and install the boto interface prior to use. To begin e-mailing workers, users must import boto, input both the access key and secret access key from their Mechanical Turk requester account, and enter both the subject of the e-mail and the message; then, individual participant IDs are entered into the terminal window in quotations with a comma between each entry. This required manually entering the IDs to ensure no error in the commas and quotations and took a significant amount of time. While using python we encountered many issues, including the software updating and therefore no longer being feasible for re-contacting participants on MTurk. Thus, we began using R software for the remainder of the data collection.

As outlined by Leeper (2014), to use R software to re-contact participants on Mechanical Turk, one must install both R software and the MTurkR R package. Within the MTurkR package, the MTurkR Wizard can be loaded. The wizard is a graphical user interface that offers a small box on the computer screen where one can select (a) single or multiple workers, (b) the subject line and body text, and (c) the number of workers being notified.

Participant characteristics

1007 participants that were recruited via Amazon.com's Mechanical Turk completed an online questionnaire. Participants had to be 18 years of age and live in the United States to be considered for participation. Three hundred sixty valid-responding

participants completed wave 4 (73% retained). There was a 31% retention rate of initial valid responders through wave 4 over the course of the entire year.

Once all data was collected, it was carefully cleaned. An excel timesheet that contained each MTurk ID and the time and date which the survey was taken, per each wave, was created and compared to an SPSS file downloaded from Qualtrics survey software. Each individual response had to be verified, as several participants incorrectly entered their individuating MTurk IDs, which then had to be manually matched to responses from prior and future waves. Any errors or repeat IDs were corrected. After this initial validation, each individual wave was examined again by another undergraduate researcher, verified once more, and organized via word, excel, and SPSS documents to ensure reliability and validity of the dataset. Both frequencies and correlations were utilized to verify reliability of all corrections. Once each wave was verified, all four waves were combined into a single file to be analyzed.

Results

To examine test-retest reliability, Spearman correlations were computed for the full scale LHA and for each of the three subscales for waves 1 and 2 ($N = 360$) and waves 3 and 4 ($N = 165$). All coefficients were significant at $p < .05$. For waves 1-2, which were completed approximately one month apart, high reliability was evidenced by scores on the LHA full scale ($r = .77$) as well as scores on the other-directed aggression ($r = .75$), self-directed aggression ($r = .76$), and antisocial behavior ($r = .70$) subscales. For waves 3-4, which were completed approximately 6 months apart, mild reliability was evidenced by scores on the LHA full scale ($r = .64$) as well as scores on the other-directed aggression ($r = .65$), self-directed aggression ($r = .64$), and antisocial behavior ($r = .59$)

For example, the Specific Loss of Interest and Pleasure Scale (SLIPS), which was validated by Winer and colleagues to examine recent changes in levels of anhedonia (also see Winer, Nadorff et al., 2014 and Winer et al., in press) was not evaluated for test-retest reliability because it is based upon the examination of the waxing and waning of symptoms of depression. Traditional conceptualizations of reliability and validity stress that an assessment can be reliable but not valid if it yields consistent scores but does not measure what it was intended to measure. However, the example of the SLIPS also shows that one can have a prospectively valid assessment that measures what it is intending to measure but has "poor reliability" when the variability of the underlying construct is not taken into account when operationalizing reliability.

Limitations

One limitation of this study is that it only contained one mode of responding: self-report. As previously mentioned, prior studies (Berman et al, 1998; Bresin et al., 2013; Coccaro et al., 1997; McCloskey and Berman, 2003; Schoenleber et al., 2011; Swogger et al., 2014; Verona et al., 2012; Victor and Klonsky, 2014) used the LHA during an interview that also collected information about life history of aggression. Solely examining self-report allowed us to collect extensive longitudinal data as part of a large overall battery, however. Thus, the efficiency of data collection and lower participant burden than what would be required for a more extensive assessment accounted for this limitation.

Another limitation of the study is the use of Amazon.com's Mechanical Turk to collect data. It is a non-clinical sample, and we did not have access to participant's medical history due to the confidential nature of the data collection and online format.

References

Berkowitz, L., 1993. Aggression: Its Causes, Consequences, and Control. McGraw-Hill, New York.

Berman, M. E., Fallon, A. E., & Coccaro, E. F. (1998). The relationship between personality psychopathology and aggressive behavior in research volunteers. *Journal Of Abnormal Psychology*, *107*(4), 651-658. doi:10.1037/0021-843X.107.4.651

Berman, M. E., McCloskey, M. S., Fanning, J. R., Schumacher, J. A., & Coccaro, E. F. (2009). Serotonin augmentation reduces response to attack in aggressive individuals. *Psychological Science*, *20*(6), 714-720. doi:10.1111/j.1467-9280.2009.02355.x

Bresin, K., Finy, M. S., & Verona, E. (2013). Childhood emotional environment and self-injurious behaviors: The moderating role of the BDNF Val66Met polymorphism. *Journal Of Affective Disorders*, *150(*2), 594-600. doi:10.1016/j.jad.2013.01.050

Brigham, J., Lessov-Schlaggar, C. N., Javitz, H. S., McElroy, M., Krasnow, R., & Swan, G. E. (2008). Reliability of adult retrospective recall of lifetime tobacco use. *Nicotine & Tobacco Research, 10,* 287-299.

Brigham, J., Lessov-Schlaggar, C. N., Javitz, H. S., McElroy, M., Krasnow, R., & Swan, G. E (2009). Test-retest reliability of web-based retrospective self-report of tobacco exposure and risk. *Journal of Medical Internet Research, 11(3).* doi: 10.2196/jmir.1248\

L., & Lemmens, P. H. H. M. (2004). Measurement of lifetime alcohol intake: utility of a self-administered questionnaire. *American Journal of Epidemiology, 159(8),* 809-817. *doi*: 10.1093/aje/kwh102

Horton, J. J., & Chilton, L. B. (2010). The labor economics of paid crowdsourcing. In *Proceedings from EC '10: The 11th ACM Con- ference on Electronic Commerce* (pp. 209–218). New York, NY: ACM. doi:10.1145/1807342.1807376

Hosie, J., Gilbert, F., Simpson, K., & Daffern, M. (2014). An examination of the relationship between personality and aggression using the general aggression and five factor models. *Aggressive Behavior*, *40*(2), 189-196. doi:10.1002/ab.21510

Kaplan, R., & Saccuzzo, D. (2013). *Psychological testing: Principles, applications, and issues*. Belmont, CA: Wadsworth Publishing.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2151-160. doi:10.1007/BF02288391

Leeper, T. J.. (2014, December 7). Introduction to the Simple Wizard (Text Based) [Web log post]. Retrieved from https://github.com/leeper/MTurkR/wiki/Wizard-Text-Based

Mueller, P., & Chandler, J. (2012). Emailing workers using Python. *Available at SSRN 2100601*.

McCloskey, M. S., & Berman, M. E. (2003). Alcohol intoxication and self-aggressive behavior. Journal Of Abnormal Psychology, *112*(2), 306-311. doi:10.1037/0021-843X.112.2.306

partner violence: The moderating role of psychopathic traits. *Criminal Justice And Behavior*, *39*(7), 910-922. doi:10.1177/0093854812438160

Swogger, M. T., Van Orden, K. A., & Conner, K. R. (2014). The relationship of outwardly directed aggression to suicidal ideation and suicide attempts across two high-risk samples. *Psychology Of Violence*, *4*(2), 184-195. doi:10.1037/a0033212

Swogger, M. T., Walsh, Z., Maisto, S. A., & Conner, K. R. (2014). Reactive and proactive aggression and suicide attempts among criminal offenders. *Criminal Justice And Behavior*, *41*(3), 337-344.

Swogger, M. T., You, S., Cashman-Brown, S., & Conner, K. R. (2011). Childhood physical abuse, aggression, and suicide attempts among criminal offenders. *Psychiatry Research*, *185*(3), 363-367. doi:10.1016/j.psychres.2010.07.036

Tavakol, M., & Dennick, R. (2011). Making sense of cronbach's alpha. International journal of medical education, 2, 53.

Tcheremissine, O. V., Lane, S. D., Lieving, L. M., Rhoades, H. M., Nouvion, S., & Cherek, D. R. (2005). Individual differences in aggressive responding to intravenous flumazenil administration in adult male parolees. *Journal Of Psychopharmacology*, *19*(6), 640-646. doi:10.1177/0269881105056532

Verona, E., Sprague, J., & Javdani, S. (2012). Gender and factor-level interactions in psychopathy: Implications for self-directed violence risk and borderline personality disorder symptoms. *Personality Disorders: Theory, Research, And Treatment*, *3*(3), 247-262. doi:10.1037/a0025945

Verona, E., Sprague, J., & Sadeh, N. (2012). Inhibitory control and negative emotional