Psychoneuroendocrinology (2014) 49, 299—309

Available online at www.sciencedirect.com

ScienceDirect

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

Variability and reliability of diurnal cortisol in younger and older adults: Implications for design decisions

Suzanne C. Segerstrom[a,*], Ian A. Boggero[a], Gregory T. Smith[a], Sandra E. Sephton[b]

[a] Department of Psychology, University of Kentucky, United States
[b] Department of Psychological and Brain Sciences, University of Louisville, United States

Received 11 July 2013; received in revised form 20 June 2014; accepted 22 July 2014

KEYWORDS
Salivary cortisol;
Generalizability theory;
Reliability;
Longitudinal design;
Individual differences

Summary The extant research is inconclusive regarding the best sampling methods to construct reliable measures of between-person differences in derived parameters of diurnal cortisol, and no study provides such recommendations for detecting within-person changes. These studies determined how many days of sampling are necessary to assess between-person differences and within-person changes over multiple occasions in diurnal mean, diurnal slope, and area under the curve (AUC). Generalizability and decision analyses were conducted on diurnal salivary cortisol data from two separate longitudinal studies, one with younger adults ($N = 124$) and one with older adults ($N = 148$). In both studies, results indicated that 3 days of data collection provided the minimal level of reliability in mean cortisol to detect between-person differences; 4—8 days were necessary to reliably assess AUC, and 10 days for cortisol slope. Similarly, in order to reliably characterize within-person changes across occasions, at least 3 days of data collection were needed for mean cortisol and AUC and 5—8 days for slope. Results also indicated that only two samples per day, taken morning and evening, could faithfully reproduce the diurnal slope calculated from 3 or 4 samples ($r = .97—.99$). Instead of having participants provide many samples per day over the course of a few days, we recommend collecting fewer samples per day over more days.

1. Introduction

Many studies explore the relationship between salivary cortisol and personal, situational, and environmental psychosocial variables. The present paper generates ''physiometric'' (Segerstrom and Smith, 2012) data to inform design

* Corresponding author at: University of Kentucky, Department of Psychology, 125 Kastle Hall, Lexington, KY 40506-0044, United States. Tel.: +1 859 257 4549; fax: +1 859 323 1979.
E-mail address: segerstrom@uky.edu (S.C. Segerstrom).

http://dx.doi.org/10.1016/j.psyneuen.2014.07.022
0306-4530/

decisions about diurnal cortisol in such studies. Measures taken on multiple people at multiple times contain three sources of variability: between-person, within-person, and measurement error. Between-person variability reflects how one person differs from another. Within-person variability reflects how people differ from themselves across time. Measurement error reflects the difference between the true state of a person at the time of measurement and the results of that measurement. We provide estimates of these sources of variability and use these estimates to predict generalizability associated with various designs. By doing so, we hope to promote (1) study design that respects the measurement properties of diurnal cortisol, (2) the regular reporting of physiometric information, and (3) further research specifically testing generalizability in varied samples and varied designs and over varied intervals. Understanding and maximizing the reliability of measures of biological variables is important to study designs that will yield accurate estimates. In a Monte Carlo analysis, when predicting an unreliable (.29) measure of immunity taken from a single occasion of measurement, 227/1000 beta weights fell outside the 95% confidence interval obtained using a reliable (.84) measure taken from an aggregation across occasions. Aggregating across more measurement occasions and thereby increasing reliability decreased the number of anomalous results (Segerstrom et al., 2006).

When assessed on consecutive days, approximately half of the variance in both cortisol level and diurnal slope is stable between-person variability, and half is idiosyncratic to the day (Kirschbaum et al., 1990; Kraemer et al., 2006; Golden et al., 2011; Kertes and van Dulmen, 2012; Ranjit et al., 2009; Ross et al., 2014; though see Hruschka et al., 2005). However, as intervals increase to weeks or months, the proportion of stable between-person variability decreases to approximately 10% (Kirschbaum et al., 1990; Hruschka et al., 2005; Ross et al., 2014; Rotenberg et al., 2012; Shirtcliff et al., 2012; c.f., Gex-Fabry et al., 2012). A person's cortisol parameters measured today have limited generalizability to other time points, even yesterday or tomorrow. The necessity of multiple assessment days in extracting stable person variance in cortisol parameters is therefore widely recognized; however, recommendations for how many days vary. The MacArthur Network's online recommendation is ''3—4 days to get a reliable assessment of a 'trait' daily concentration (area-under-the-curve), and 6 or more days to get a reliable assessment of a 'trait' rhythm.'' Kraemer and colleagues (2006) estimated 2—3 days to reliably estimate ''trait'' slope. Hruschka and colleagues (2005) estimated 14—22 days to reliably estimate ''trait'' slope (albeit fewer for cortisol level). Clearly, ''more studies need to be carried out ... to define the precise parameters of sampling requirements'' (Goodyer et al., 2001, p. 243).

In studies with multiple measurement occasions (e.g., an intervention study with baseline, post-intervention, and follow-up or a longitudinal study with annual measurement occasions), cortisol might be measured on multiple days within each occasion. Variance estimates for single-day measurements are of limited utility for design decisions involving multiple measurements at multiple time points (Kirschbaum et al., 1990; Hruschka et al., 2005; Rotenberg et al., 2012; Shirtcliff et al., 2012). Instead, decisions may focus on the number of days and occasions required for measures to discriminate people from each other at the same occasion, different occasions, or across an aggregate of occasions with adequate reliability. Furthermore, there is the question of how many days per occasion would be required for a measure to discriminate a person at one occasion from himself or herself at a different occasion with adequate reliability (e.g., before and after treatment) (Cranford et al., 2006). Table 1 provides examples of study designs that focus on discriminating people from each other when measured once at the same occasion (I), when measured once at different occasions (II), and when measured across several occasions (III), as well as a study design that focuses on discriminating change within people across several occasions (IV).

The challenge in designing diurnal cortisol studies is to maximize variability arising from the facet of interest (e.g., differences between or changes within people) and minimize variability due to other facets and measurement error. Classical test theory assumes that any observed score is the result of a true score and error variance. Generalizability theory (Shavelson and Webb, 1991; Brennan, 2001) extends this assumption to encompass multiple sources of variance, for example, due to people, occasions, and their interaction. What variance is of interest and therefore ''true score'', however, depends on the design and research question. For example, to demonstrate that ''individuals differ in their patterns of cortisol secretion and ... these differences exhibit some stability over time'' (Hruschka et al., 2005, p. 699), the ''true score'' of a cortisol parameter would consist of person variance, and ''error'' variance would arise from day and occasion variance, as well as their interactions (Cranford et al., 2006). In other cases, the variance across occasions (i.e., change over time) may be the ''true score'', and ''error'' variance would arise from variance among the days comprising each occasion. (Measurement error per se typically accounts for very little variability in cortisol when assays are done competently; Kirschbaum et al., 1990; Kertes and van Dulmen, 2012; Marceau et al., 2013). Even the variability of scores can be the ''true score'' of interest (e.g., Marceau et al., 2013), but it is still necessary to provide for adequate reliability in the measurement of variability, itself a methodological challenge (Estabrook et al., 2012).

Finally, derived measures such as diurnal slope require consideration of another design decision involving the number of samples collected each day and their timing (Kudielka et al., 2012). Some studies of diurnal slope have asked participants to provide over 40 samples (Ice et al., 2004), but other evidence suggests two to three samples per day may effectively reproduce the slope calculated from more samples (Kraemer et al., 2006). Collecting many samples per day is expensive for researchers and burdensome to participants. Therefore, further research is needed to clarify how well fewer samples reproduce slopes calculated using more samples per day.

2. The current studies

We applied generalizability theory (Shavelson and Webb, 1991; Brennan, 2001) to two longitudinal studies, one with younger adults and one with older adults. Generalizability