International Journal of Behavioral Medicine (2020) 27:262–266
https://doi.org/10.1007/s12529-020-09899-0

SPECIAL ISSUE: SALIVARY BIOSCIENCE

Physiometrics in Salivary Bioscience

Suzanne C. Segerstrom [1]

Published online: 13 May 2020

Abstract

Background Accurate estimation in statistical models depends on sample size but also, critically, reliability of the measure. Physiometrics is the equivalent of psychometrics for measures such as sex hormones, catabolic hormones, and products of the immune system.

Method There are multiple ways to measure physiometrics, from simple correlation to complex generalizability theory designs. Depending on the design, these estimates can provide information about equivalency (e.g., the correlation between two measurements taken close together in time) or stability (e.g., the correlation between two measurements taken farther apart in time).

Results The physiometrics of salivary measures including cortisol, α-amylase, testosterone, and cytokines range from highly stable, requiring only a single sample, to highly unstable, requiring multiple samples to achieve generalizability to longer periods of time. However, generalizability is relative to the study design, and only some designs call for stable and generalizable measures.

Conclusion Both dedicated physiometric studies and more reporting of physiometrics in psychoneuroendocrinology and psychoneuroimmunology will improve the quality of salivary bioscience study designs in the future.

Keywords Cortisol · Testosterone · Cytokine · Reliability · Generalizability

Introduction

> Variability introduces error into behavioral studies, where stable measures are needed to characterize individual differences and changes over time. Without more information on this variability, one cannot know how many subjects to run, how many measurements to take, and when to take the measurements [1, p. 83].

Although this epigraph could apply to any number of measures, in this case, Dabbs [1] refers to the variability of salivary testosterone (T). The idea of "physiometrics" [2, 3]—the physiological equivalent of psychometrics—seems to have been around for at least 3 decades. Although the validity of physiological measures, especially salivary measures, is a major research topic (e.g., the correlation between salivary and serum levels), there is less attention to variability, generalizability, and measure reliability (vs. assay reliability [4]). Here, I hope to convince readers that there are good reasons that more attention to physiometrics will benefit salivary bioscience.

Most important, accurate estimation depends on larger sample sizes and more reliable measurement; however, it is possible to trade off between the two [5–10]. The issue of "how many subjects to run, how many measurements to take" [1, p. 83] reflects consideration of this trade-off. When samples are small and measures are unreliable, estimates in statistical models are often substantially too large or small and/or reflect the wrong pattern of results (e.g., a beta weight in the opposite direction from the true effect). One of the statistical guidelines for *International Journal of Behavioral Medicine* is report reliability "(1) for the analytic sample, (2) for all measures including biomarkers, and (3) taking into consideration the design of the study" [11, p. 456]. A statistical model, unfortunately, does not care how difficult a sample is to recruit or how expensive assays are. Researchers, reviewers, and editors should all consider whether a sample size is sufficient to compensate for a measure with low reliability (as might be true in a very large, population-based survey) or whether reliability is sufficient to compensate for a small sample size (as might be true when studying a rare condition or demographic).

✉ Suzanne C. Segerstrom
segerstrom@uky.edu

[1] Department of Psychology, University of Kentucky, 125 Kastle Hall, Lexington, KY 40506-0044, USA

Quantifying Physiometrics

This section gives a brief overview of ways to measure physiometrics; for more detail on these approaches, see reference [12]. All these approaches are based on covariances or correlations among observations. As such, all of the same problems that distort or attenuate a correlation (e.g., non-normal distributions, restriction of range) can distort physiometrics. Right-skewed distributions are characteristic of many physiological measures, and it is a common practice to log transform them to normalize the distribution before analysis. This practice should also be employed when calculating physiometrics.

For two occasions of measurement, correlations between them can take two forms: when taken at different times, the correlation can reflect stability (i.e., test-retest), or when taken at the same time, equivalence (i.e., parallel forms). For any study design, what constitutes the same or different times may vary. For a very short-term study, equivalence measures might be taken one after the other, and stability measures might be taken 2 days in a row. For a very long-term study, equivalence measures might be taken 2 weeks in a row, and stability measures might be taken over an interval of several years.

For more than two measurements, equivalence and stability can be quantified several ways, but all of them ask the same basic question: Of all the variability in a measure, how much of it can be attributed to the variance of interest? Cronbach's coefficient alpha assumes that each observation (e.g., item on a scale or physiological measurement) covaries equally with the true score. Given the average covariance and the number of observations and applying the Spearman-Brown formula, the resulting coefficient should be the square of the correlation between the true score and the obtained score (i.e., the percent of true score variance). As the number of observations increases, reliability will also increase, as implied by the Spearman-Brown formula.[1] Decreased error in observations (e.g., better assay reliability) can also increase alpha, assuming that it increases true score variance in the observed scores. Notably, although alpha is not the best measure of reliability for most psychological measures because of the assumption of equal covariances, it is an appropriate measure for repeatedly measured physiological measures for which that assumption is reasonable. The span of time over which the measures are taken can suggest equivalence or stability, and equivalence or stability over 3 days does not imply equivalence or stability over 3 months.

Cronbach later developed generalizability theory, in which there can be more than one facet of "true score": For example, in longitudinal burst designs (intensive data collections repeated over a longer period of time), there can be a true score for the person across all bursts and another true score within each burst. A generalizability study estimates the amount of variance due to each facet and the interactions among them, and a decision study estimates how many measurements are needed to reliably capture those sources of variance. For example, for salivary cortisol diurnal slope, 11% of the variance was due to stable individual differences, 14% of the variance was due to person by occasion interactions (people reacting differently to changing circumstances), and 75% of the variance was due to idiosyncratic cortisol slopes on a specific day, not systematically related to person, occasion, day, or their two-way interactions. Consequently, reliable (> 0.80) measurement of stable individual differences in a longitudinal burst design could be achieved with 10 days of sampling over 3 occasions or 3 days of sampling over 10 occasions [13].

The intraclass correlation (ICC), like alpha and the generalizability coefficient, is the ratio of the variance of interest over total variance and likewise indicates the percent of true score variance associated with the class of interest. It can be interpreted as the correlation between any two members of a class (for example, a class might be a person and ten observations on that person, the members of that class). The difference between the ICC and other measures of reliability based on classical test theory is that absolute levels in a measure, not just relative levels, contribute to in the denominator. Therefore, the ICC is more conservative than alpha or the generalizability coefficient. However, generalizability theory does have the ability to consider absolute levels in reliability with the dependability index. Like the ICC, a dependability index includes variance due to absolute levels in the denominator, whereas alpha and the generalizability coefficient do not.

Physiometrics in Salivary Bioscience

There are different amounts of evidence available, but physiometrics have been reported for several salivary measures. The most evidence is available for salivary cortisol measures, typically the diurnal slope, diurnal area under the curve (AUC), and/or awakening response (CAR). Equivalence ICCs across days for diurnal slope and CAR were lowest, between 0.00 and 0.20; those for AUC were slightly higher, between 0.00 and 0.30. These estimates agree with a generalizability study finding person variance to be ~ 11% for slope and 10–20% for AUC [13]. Stability ICCs across months to years for cortisol measures were higher,

[1] The Spearman-Brown formula, where ρ is, for the purposes of alpha, the average covariance between observations (x and x') and n is the number of observations

$$\rho^{*}_{xx'} = \frac{n \times {}^{*}\rho_{xx'}}{1 + (n-1) \times {}^{*}\rho_{xx'}}$$

Table 1 Stability coefficients for salivary and serum cytokines and CRP

Cytokine	Salivary *r* (18 months) [19]	Salivary *r* (1–2 years) [20]	Typical serum ICC (months to years) [2]
Tumor necrosis factor α	0.22	0.18–0.35	~0.40–0.90
IL-6	0.10	0.19–0.30	~0.50
IL-8	0.27	0.28–0.45	~0.40–0.70
CRP	0.31	–	~0.50–0.70

ICC intraclass correlation, *IL* interleukin, *CRP* C-reactive protein

typically 0.20–0.45 for slope, 0.50–0.75 for AUC, and 0.10–0.20 for the CAR. Longer intervals generally produced lower ICCs (see [2] for a review of salivary cortisol physiometrics).

Salivary α-amylase (sAA) has not been studied as extensively as salivary cortisol, and so there are only two studies to my knowledge that reported its physiometrics. Over 24 months, the stability ICC was high (0.75) for diurnal AUC but lower (0.43) for the awakening response [14]. These estimates are comparable to those for person variance in a 6-month generalizability study (0.61 for AUC and 0.26 for awakening response [15]).

Both equivalency and stability correlations have been reported for T. Over 2 days, the typical equivalency was $r = 0.64$ for both men and women. Stability declined from $r = 0.71$ (1–2 weeks) to $r = 0.52$ (7–8 weeks) in men only [1]. These estimates agree with stability over 2 weeks reported for men ($r = 0.78$) and women ($r = 0.65$) [16]. In the latter study, progesterone stability was $r = 0.32$ for both men and women. These estimates for the stability of T lie between two differing ICCs reported for serum T (0.92 over 4 weeks in post-menopausal women [17] and 0.31 over 3 months in adult men [18]). A single study reporting physiometrics of both salivary and serum T in men and women would be very useful in understanding how much equivalency and stability can be expected, as well as the role of sex and age in promoting or suppressing variability.

Salivary cytokines vary in their equivalency and stability. A large sample of adolescent girls had saliva sampled twice on the same day (equivalency), repeated after 18 months (stability). Equivalency *r* values ranged between 0.51 (interleukin (IL)-8) and 0.81 (C-reactive protein (CRP)) [19]. Most analytes could achieve good equivalency or reliability with 1–2 samples measured at the same time point, with the exception of IL-18. However, the stability of the mean of the two samples from each time point was lower, ranging from 0.10 (IL-6) to 0.37 (IL-18). Another sample of adolescent girls yielded similar stability estimates over 1 year (mean *r* across cytokines = 0.25–0.30) and lower estimates than the same cytokines measured in serum ($r = 0.33$–0.61) [20]. Both studies' estimates were markedly lower than ICCs reported for serum cytokines measured over periods of months to years. Table 1 shows the salivary and serum estimates (see [2] for a review of serum cytokine physiometrics). These salivary cytokine studies sampled a very specific population, and other or broader populations may yield different results. For example, later pubertal stage was associated with lower cytokine concentrations [20], suggesting that stability might have compromised by individual differences in maturation.

Physiometrics in the Special Issue

Perhaps the most important thing to remember is that good generalizability for one study design does not imply good generalizability for another design. In generalizability theory, one has to define a "universe of generalization": for example, for what people over what period of time do you want to generalize?

For some study designs, the issue of generalizability may not be relevant. In an experimental study, you do not need to generalize beyond the moment of the study if you are only interested in appraisals during the experiment, for example [21; this issue]. However, you might want to generalize one measurement of experimental reactivity to a longer-term exposure [22; this issue]. The stability of reactivity is itself a topic of investigation [2]. Similarly, you might want to generalize from a few days' measurement only to those days. For example, if diurnal cortisol is being predicted from that day's (or the previous day's) social support [23; this issue], you do not need the cortisol measurement to generalize beyond those days.

For other study designs, the issue of generalizability and the universe of generalization are very relevant. For example, psychological states over a week might be used to predict a salivary biomarker at the end of that week [24; this issue]. You want the biomarker measurement to generalize to the previous week. If it did not generalize past the day on which it was collected (as is true for some salivary cortisol measures), it would not be useful. You would need to collect saliva throughout the week to make the hypothesized inference. In this case, the biomarkers were salivary CRP and IL-6, which have poor generalizability across a period of months to years, but that is not the universe of generalization for this study. In the absence of appropriate physiometrics, you could either hope that physiometrics of serum biomarkers generalize to salivary biomarkers (cross your fingers!) or design your study blindly.

The former is better than the latter, but neither is ideal. More knowledge of physiometrics in this domain will help scientists make better design decisions.

If you are examining a variable biomarker, it behooves you to consider this further wisdom from Dabbs [1, p. 85]: "Scores that are farther apart in time are less likely to be affected by the same transient events. Scores farther apart in time should thus more effectively cancel momentary highs and low to provide a true picture of subjects' characteristic testosterone [or other biomarker] concentrations." If you are interested in cortisol exposure during pregnancy, repeated measures across pregnancy are the right way to go. Gestation is your universe of generalization [25; this issue]. On the other hand, you might be examining a more stable biomarker such as T. In that case, a cross-sectional design might yield results that could generalize to the surrounding weeks or months [26; this issue].

Conclusion

Salivary biomarkers range from highly variable (cortisol) to highly stable (T) and in between (cytokines). We do not know as much as we should about the physiometrics of most salivary biomarkers. However, we can use what we know about their physiometrics to better define our universes of generalization (moment, day, week, trait) and ensure that we have sufficient reliability and generalizability (or a large enough N to compensate for reliability that is lower than we would like [6–10]). This special issue includes study designs with varying universes of generalizability and demonstrates how one can use different biomarkers with different physiometrics to make interesting inferences.

This article does not contain any studies with human participants or animals performed by any of the authors.

Acknowledgments The author thanks Michael A. Hoyt for his helpful comments on previous versions.

Compliance with Ethical Standards

Conflict of Interest The author declares that she has no conflict of interest.

Informed Consent There were no participants in this review, and therefore, no informed consent was necessary.

Ethical Approval For this type of review, ethical approval is not required.

References

1. Dabbs JM. Salivary testosterone measurements: reliability across hours, days, and weeks. Physiol Behav. 1990;48:83–6.
2. Gloger EM, Smith GE, Segerstrom SC. Stress physiology and physiometrics. In: Ragin DF, Keenan JP, editors. Handbook of research methods in health psychology. New York: Routledge. In press.
3. Segerstrom SC, Smith G. Methods, variance, and error in psychoneuroimmunology research: the good, the bad, and the ugly. In: Segerstrom SC, editor. Oxford handbook of psychoneuroimmunology. New York: Oxford; 2012. p. 421–32.
4. Schultheiss OC, Stanton SJ. Assessment of salivary hormones. In: Harmon-Jones E, Beer JS, editors. Methods in social neuroscience. New York: Guilford; 2009. p. 17–44.
5. Gelman A, Carlin J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Persp Psychol Sci. 2014;9:641–51.
6. Kraemer HC. To increase power in randomized clinical trials without increasing sample size. Psychopharmacol Bull. 1991;27:217–24.
7. Leon AC, Marzuk PM, Laura P. More reliable outcome measures can reduce sample size requirements. Arch Gen Psychiatry. 1995;52:867–71.
8. Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. Biol Psychiatry. 2000;47:762–6.
9. Segerstrom SC. Between the error bars: how modern theory, design, and methodology enrich the personality-health tradition. Psychosom Med. 2019;81:408–14.
10. Segerstrom SC, Boggero IA. The emperor has no cortisol: frequency of estimation errors in biomarker studies. https://vimeo.com/399224361. Accessed 17 April 2020.
11. Segerstrom SC. Statistical guideline #2: report appropriate reliability for your sample, measure, and design. Int J Behav Med. 2019;26:455–6.
12. Webb NM, Shavelson RJ, Haertel EH. Reliability coefficients and generalizability theory. In: Rao CR, Sinhara S, editors. Handbook of statistics 26: psychometrics. New York: Elsevier; 2006. p. 81–124.
13. Segerstrom SC, Boggero IA, Smith GT, Sephton SE. Variability and reliability of diurnal cortisol in younger and older adults: implications for design decisions. Psychoneuroendocrinol. 2014;49:299–309.
14. Skoluda N, La Marca R, Gollwitzer M, et al. Long-term stability of diurnal salivary cortisol and alpha-amylase secretion patterns. Physiol Behav. 2017;175:1–8.
15. Out D, Granger DA, Sephton SE, Segerstrom SC. Disentangling sources of individual differences in diurnal salivary α-amylase: reliability, stability and sensitivity to context. Psychoneuroendocrinol. 2013;38:367–75.
16. Liening SH, Stanton SJ, Saini EK, Schultheiss OC. Salivary testosterone, cortisol, and progesterone: two-week stability, interhormone correlations, and effects of time of day, menstrual cycle, and oral contraceptive use on steroid hormone levels. Physiol Behav. 2010;99:8–16.
17. Cauley JA, Gutai JP, Kuller LH, Powell JG. Reliability and interrelations among serum sex hormones in postmenopausal women. Am J Epidemiol. 1991;133:50–7.
18. Brambilla DJ, O'Donnell AB, Matsumoto AM, McKinlay JB. Intraindividual variation in levels of serum testosterone and other reproductive and adrenal hormones in men. Clin Endocrinol. 2007;67:853–62.
19. Shields GS, Slavich GM, Perlman G, Klein DN, Kotov R. The short-term reliability and long-term stability of salivary immune markers. Brain Behav Immun. 2019;81:650–4.
20. Riis JL, Out D, Dorn LD, Beal SJ, Denson LA, Pabst S, et al. Salivary cytokines in healthy adolescent girls: Intercorrelations, stability, and associations with serum cytokines, age, and pubertal stage. Dev Psychobiol. 2014;56:797–811.

21. Shirotsuki K, Izawa S, Sugaya N, Kimura K, Ogawa N, Yamada KC, et al. Imbalance between salivary cortisol and DHEA responses is associated with social cost and self-perception to social evaluative threat in Japanese healthy young adults. Int J Behav Med. In press.
22. Eiden RD, Shisler S, Granger DA, Schuetze P, Colangelo J, Huestis MA. Prenatal tobacco & cannabis exposure: associations with cortisol reactivity in early school age children. Int J Behav Med. In press.
23. Hooker ED, Campos B, Hoffman L, Zoccola P, Dickerson SS. Is receiving social support costly for those higher in subjective socioeconomic status? Int J Behav Med. In press.
24. Slavish DC, Jones DR, Smyth JM, Engeland CG, Song S, McCormick NM, et al. Positive and negative affect and salivary measures of inflammation among young adults. Int J Behav Med. In press.
25. Riis JL, Granger DA, Woo H, Voegtline K, DiPietro JA, Johnson SB. Long-term associations between prenatal maternal cortisol and child neuroendocrine-immune regulation. Int J Behav Med. In press.
26. Martin LA, Ter-Petrosyan M. Positive affect moderates the relationship between salivary testosterone and a health behavior composite in university females. Int J Behav Med. In press.