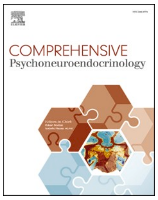
Contents lists available at ScienceDirect

Comprehensive Psychoneuroendocrinology

journal homepage: www.sciencedirect.com/journal/comprehensive-psychoneuroendocrinology



Reliability of diurnal salivary cortisol metrics: A meta-analysis and investigation in two independent samples

Sara A. Norton [a,*], David AA. Baranger [a], Ethan S. Young [b], Michaela Voss [a], Isabella Hansen [a], Erin Bondy [a], Merlyn Rodrigues [a], Sarah E. Paul [a], Elizabeth Edershile [c], Patrick L. Hill [a], Thomas F. Oltmanns [a], Jeffry Simpson [c], Ryan Bogdan [a]

[a] *Washington University in St. Louis, Department of Psychological & Brain Sciences, USA*
[b] *Utrecht University, Department of Psychology, the Netherlands*
[c] *University of Minnesota, Department of Psychology, USA*

ARTICLE INFO

Keywords:
Cortisol
Reliability
Stress
Variability

ABSTRACT

Stress-induced dysregulation of diurnal cortisol is a cornerstone of stress-disease theories; however, observed associations between cortisol, stress, and health have been inconsistent. The reliability of diurnal cortisol features may contribute to these equivocal findings. Our meta-analysis (5 diurnal features from 11 studies; total participant n = 3307) and investigation (15 diurnal cortisol features) in 2 independent studies (St. Louis Personality and Aging Network [SPAN] Study, n = 147, ages 61–73; Minnesota Longitudinal Study of Risk and Adaptation [MLSRA] Study, n = 90, age 37) revealed large variability in the day-to-day test-retest reliability of diurnal features derived from salivary cortisol data (i.e., ICC = 0.00–0.75). Collectively, these data indicate that some commonly used diurnal cortisol features have poor reliability that is insufficient for individual differences research (e.g., cortisol awakening response) while others (e.g., area under the curve with respect to ground) have fair-to-good reliability that could support reliable identification of associations in well-powered studies.

1. Introduction

All organisms seek to maintain homeostasis by dynamically regulating physiology and behavior in response to internal and external cues. Stress, which occurs when one perceives inadequate resources to confront environmental pressures, disrupts homeostasis [1]. Although stress can evoke adaptive responses to immediate challenges when motivation is high and resources are available, stress exposure is among the strongest predictors of the onset and course of many negative mental and physical health outcomes [2]. The ubiquity of stress and its pathogenic effects has inspired extensive efforts to understand the biological mechanisms through which stress may influence health.

Studies evaluating biological mechanisms through which the pathogenic effects of stress emerge predominantly focus on cortisol [3]. As the end product of the neuroendocrine hypothalamic-pituitary-adrenal (HPA) axis, cortisol follows a diurnal rhythm wherein levels peak immediately upon awakening and decline slowly over the course of the day. Stress exposure may induce negative health outcomes by dysregulating this diurnal rhythm [4]. Specifically, evidence has linked blunted (i.e., flatter) diurnal cortisol fluctuation to negative health outcomes [5] and shown that such blunted diurnal rhythm can be induced by chronic stress in non-human animal models [6] and is associated with chronic stress in humans [7]. This evidence, combined with practicalities (e.g., ease of at-home collection and quantification via reliable inexpensive assays), has contributed to cortisol beings one of the most studied biological phenotypes of stress and health.

Unfortunately, though pervasive in stress- and health-related research, cortisol studies tend to produce mixed results. For example, an abnormal cortisol awakening response (CAR) has been associated with depression in some studies (e.g. Refs. [8,9], but results are inconsistent (e.g., Refs. [10,11]. Among many factors (e.g., study design differences), widespread variability in reliability (i.e., measurement/estimation consistency) may contribute to these inconsistent findings. As the observable correlation between two variables is their true correlation multiplied by the square root product of their reliability, poor measurement reliability lowers statistical power [12,13] and can produce misleading results. State-related variability and measurement error in diurnal cortisol features has been well

* Corresponding author.
E-mail address: s.norton@wustl.edu (S.A. Norton).

https://doi.org/10.1016/j.cpnec.2023.100191
Received 6 July 2023; Accepted 6 July 2023
Available online 9 August 2023

2666-4976/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

acknowledged [14,15]; however, only a few studies have assessed the reliability of day-to-day diurnal cortisol metrics, with reliability estimates ranging from 0.15 to 0.78 (Table 1). Thus, a critical step towards producing more consistent results is systematically examining the reliability of common diurnal cortisol metrics.

In the present study, we *first* conducted a meta-analysis of existing day-to-day diurnal cortisol reliability studies (n = 11 studies from 10 publications). *Second*, we estimated the reliability of day-to-day diurnal cortisol features (e.g., cortisol awakening response, area under the curve, etc.) in two independent samples: the Saint Louis Personality and Aging Network (SPAN) study (n = 147, ages 61–73) and the Minnesota Longitudinal Study of Risk and Adaptation (MLSRA; n = 90, age 37). Using data from SPAN, we examined whether protocol adherence and other factors (childhood maltreatment, race, sex) influenced cortisol reliability. Given the burden of saliva collection on participants and cortisol processing on researchers, we also estimated whether the most reliable index of diurnal cortisol could be recapitulated using fewer samples per day and whether such an index is reliable. We expected wide variability in estimated reliability across diurnal cortisol features. More specifically, we hypothesized that features that integrate more cortisol samples (e.g. area under the curve metrics) would be more reliable than difference scores (e.g. cortisol awakening response). This would be consistent with past reliability studies (Table 1) as well as evidence that the combination of multiple measures typically improves reliability [16] while difference scores lower measurement reliability [17].

2. Methods

2.1. *Meta-analytic reliability of cortisol features*

A systematic search was used to identify published studies reporting on the reliability of salivary diurnal cortisol metrics. PubMed and Google Scholar as well as the reference lists of included studies were searched by author SAN. Search terms included “cortisol” AND “reliability” OR “stability” OR “ICC.” Only peer-reviewed empirical journal articles published in English and reporting on ICC of salivary cortisol across at least 2 days and within 30 days were included. Dates of publication and age of subjects were not restricted, but studies were excluded if there was not a healthy control group. Our search identified 10 published manuscripts reporting on diurnal features of cortisol reliability for meta-analysis (Table 1). To calculate confidence intervals around reported ICCs, a Fisher *r*-to-Z transformation was applied, confidence intervals were calculated, and the results were back-transformed. Then, a random-effects meta-analytic model was fitted using the *metafor* package in R to report meta-analytic ICCs [18].

2.2. *Independent studies assessing diurnal cortisol reliability*

2.2.1. *Participants*

SPAN. The St. Louis Personality and Aging (SPAN) study began in 2007 and is an ongoing longitudinal protocol assessing a wide range of personality, health, social, and biological characteristics in a representative community sample of older adults residing in the St. Louis, Missouri area [19]. Individuals were excluded if they lacked a permanent residence, could not read at a 6th-grade level, or had active psychotic symptoms. Each participant completed a 3-h in-person assessment at baseline (N = 1630; baseline), and 3 subsequent in-person follow-up (IPFU) sessions occurring approximately every 2–3.5 years (see Supplemental Fig. 1). Participants were also asked to complete a short sequence of mailed or online follow-up (FU) questionnaires every 6 months after entering the study.

Take-home saliva collection kits were provided at the second IPFU session (IPFU-2; data collected between December 2014 and August 2017) to a subset of white and black participants who reported either high ($n = 73$) or low ($n = 73$; groups recruited to be balanced on race and sex) maltreatment during childhood (Table 2).[1] Childhood maltreatment was assessed at IPFU-2 using the 28-item Childhood Trauma Questionnaire (CTQ [20], which retrospectively assesses emotional, physical, and sexual abuse, as well as emotional and physical neglect, before the age of 17. Participants reporting moderate or greater levels of childhood maltreatment according to CTQ scoring cut offs on the emotional abuse (≥13), physical abuse (≥10), physical neglect (≥10), or sexual abuse (≥8) subscales were recruited for the high childhood maltreatment group (emotional neglect was not used to define the high maltreatment group). Participants in the low childhood maltreatment group reported no/minimal childhood maltreatment on all 5 CTQ subscales (≤8 for emotional abuse, ≤7 for physical abuse, ≤5 for sexual abuse, ≤9 for emotional neglect, ≤7 for physical neglect). All participants consented to the SPAN protocol, which was approved by the Washington University in St. Louis Institutional Review Board. They received $60 for each in-person session and $40 for competing salivary data collection.

MLSRA. The Minnesota Longitudinal Study of Risk and Adaptation (MLSRA) is an ongoing longitudinal study focusing on the assessment of social relationship experiences [21,22]. From 1975 to 1976, pregnant women (n = 267) living below the poverty line who were receiving free healthcare services at the time of enrollment were recruited. Their children (n = 267) became the MLSRA cohort. They were assessed at birth and every 6 months until 2 ½, then yearly through 3rd grade, three times between ages 9 and 13, and at ages 16, 17 ½, 19, 23, 26, 28, and 37. Salivary cortisol data was collected at the 37-year follow-up assessment (n = 90).

2.2.2. *Measures*

2.2.2.1. *Salivary Cortisol Collection and Assay.* ***Cortisol Collection in SPAN.*** Participants were issued take-home saliva collection kits that consisted of instructions (Supplementary Material), a self-report sample log, and a jug fitted with a Medication Event Monitoring System (MEMS®) Cap (Aardex Group, Serain, Belgium) that logged each time the jug was opened. Inside the jug were 28 Salivettes® (Sarstedt) to facilitate the collection of 6 saliva samples/day over 4 days (only 3 days were assayed, described below) as well as 4 extra Salivettes that could be used if collection issues occurred.

Participants were instructed to collect samples on sequential weekdays at the following times: immediately upon waking (before getting out of bed or doing any activity, **T1**), 30 min after waking (**T2**), 2.5 h after waking (**T3**), 8 h after waking (**T4**), 12 h after waking (**T5**), and bedtime (**T6**). Participants were asked to record the exact times of sample collection in a log book and told that the MEMS Cap would also log each time the jug was opened. Self-reported and MEMS® Cap times were highly correlated (r = 0.94; additional information provided in Supplemental Table 1). Self-reported time was used for all analyses but was replaced with MEMS® Cap times when self-reported times were missing (n = 21). Participants were also asked to record instances of food, caffeine, alcohol, and tobacco intake, exercise, stressful events they may have experienced, and medications or drugs they took that day.

Cortisol Assay in SPAN. For each participant, 3 of the 4 collected days were assayed (total n = 2646 samples assayed). All within participant samples were collected within 4 days of one another. The non-assayed day was selected for exclusion due to relatively worse adherence (e.g., sample collected at a time different from instructions) or the occurrence of unusual events (e.g., a major stressor). As participant data were collected within a broad temporal window (December 2014–August 2017), we explored whether data differed in reliability across 9-month windows. There was no evidence that reliability differed

[1] One person included in the analysis was not in either the “high” or “low” CTQ group, bringing the total number of participants to 147.

Table 1
Summary of studies evaluating the ICC of salivary cortisol-derived features.

	Almeida et al.	Bakusic et al.	Golden et al.	Hellhammer et al.	Kuhlman et al.	Rotenburg et al.	Tomarken et al.	Viardot et al.	Wang et al.	Zhang et al.
Year of publication	2010	2019	2014	2007	2019	2012	2015	2005	2014	2017
N	1143	18	935	193	59	264	27°	20	580	95
Mean age (age range) of sample	57 (33–84)	(23–39)	65 (45–84)	Younger group: 36.2 (26–46) Older group: 71.0 (63–88)	11.02 (8–13)	12.4 (9–18)	9.72 (7.02–12.85)	32.5 (20–58)	63.7 (45–84)	18.72 (17–21)
Samples per day (number of sampling days)	4 (4)	3 (7)	6 (3)	4 (6)	4 (8)	5 (3) (Study 1) 6 (2) (Study 2)	4 (3)	2 (1)	6 (3) (Wave 1) 8 (2) (Wave 2)	5 (3)
Time interval	4 days	1 week	3 days	6 days	3 weeks	2–64 days (97% of samples returned within 14 days)	3 days	1 day	3 days (Wave 1) 2 days (Wave 2)	2 weeks
Assay	CLIA	LC-MS/MS	CLIA	DELFIA	ELISA	DELFIA	RIA	RIA	CLIA	LC-MS/MS
Reliability Scores										
Wakeup (T1)	0.22	0.15	0.48		0.49	0.40	0.45	0.47	0.52 (Wave 1) 0.52 (Wave 2)	0.43
Peak (T2)		0.37				0.54	0.63			0.49
Late Morning (T3)						0.37				0.20
Afternoon (T4)							0.32			0.42
Evening (T5)						0.41				
Bedtime (T6)					0.50	0.21	0.44	0.78		
CAR Incline			0.28		0.46		0.38		0.31 (Wave 1) 0.18 (Wave 2)	
CAR AUC		AUCg: .25 AUCi: .29		AUCg: .85 AUCi: .71		0.49				
AUCg			0.66			0.58			0.67 (Wave 1) 0.74 (Wave 2)	0.56
AUCi						0.26				
Early Decline (ED)			0.37						0.33 (Wave 1) 0.27 (Wave 2)	
Late Decline (LD)			0.32						0.29 (Wave 1) 0.27 (Wave 2)	
Maximum Decline (MD)							0.33			
Diurnal Slope					0.71	0.27				

Table 2
Demographic information of SPAN participants included in the salivary cortisol analysis.

	Total	High Adversity	Low Adversity
Mean age (SD)	67.3 (3.1)	67.3 (3.0)	67.2 (3.2)
Age range	61–73	61–73	61–73
Sex			
Male	65 (44%)	29 (40%)	37 (51%)
Female	82 (56%)	44 (60%)	36 (50%)
Race			
Black	61 (42%)	30 (41%)	31 (43%)
White	86 (59%)	43 (59%)	42 (58%)
Education			
Less than high school	2 (1%)	2 (3%)	0 (0%)
High school or GED	20 (14%)	13 (18%)	7 (10%)
Some college	29 (20%)	13 (18%)	15 (21%)
Vocational school	10 (7%)	6 (8%)	4 (5%)
2-year degree	10 (7%)	7 (10%)	3 (4%)
4-year degree	33 (23%)	11 (15%)	22 (30%)
Master's degree	31 (21%)	16 (22%)	15 (21%)
Doctoral degree (PhD)	6 (4%)	3 (4%)	3 (4%)
Professional degree (e.g. MD or JD)	5 (3%	1 (1%)	4 (5%)

according to when samples were collected (see Supplemental Table 2).

Cortisol was assayed in duplicate using commercially available enzyme-linked immunosorbent assays (SLV2390R Salivary Cortisol ELISA DRG International Inc., USA). Average intra- and inter-assay CVs were acceptable (<9% and <15%, respectively). Samples producing unreliable measures (i.e., intra-assay CVs >20%) even after being re-assayed in duplicate were excluded (n = 46). Further, because T2 is meant to capture the peak of the cortisol awakening response, this sample was excluded if the time between T1 and T2 was ≤20 min, or ≥50 min, or if the time was not recorded (n = 42 samples).

As expected, cortisol values were positively skewed (skew = 1.62). To reduce skew and maintain consistency with prior investigations of salivary cortisol, all data were log-transformed prior to analyses (see Supplemental Table 3). All reported cortisol data represent log-transformed cortisol concentration in ng/mL. Outliers were calculated by computing means and standard deviation for each time point using the data from all three days; values that were ±2.5 standard deviations from the mean (after log-transformation) for that time point were winsorized (n = 50). None of the samples included in analyses were under the minimum detection limits of the assay. Following quality control, 2302 individual time points from 147 individuals were used in the final analytic sample. More information about removed data can be found in Supplemental Table 4.

Cortisol Collection and Assay in MLSRA. Cortisol collection and processing for the MLSRA study have been previously described [22]. Briefly, at age 37, participants ($n = 90$) provided five saliva samples on each of two consecutive days by passively drooling through a straw into a labeled vial. Participants were instructed to collect samples at the following times: immediately upon waking (**T1**), 30 min after waking (**T2**), 1 h after waking (**T3**), in the afternoon (**T4**), and just before going to bed (**T5**). MEMS® Caps were used to confirm when the saliva samples were provided and to corroborate self-reported sample times. When the self-reported and MEMS® Cap times differed, the MEMS® Cap time was used. Participants mailed their samples back to the University of Minnesota, where the samples were stored at −20 °C before being shipped to the University of Trier, Germany, for assaying using time-resolved fluorescence immunoassay (dissociation-enhanced lanthanide fluorescent immunoassay, or DELFIA). Each sample was assayed in duplicate, and results of the two assays were averaged. All cortisol data were log-transformed prior to analyses to correct for positive skew. Log-transformed data were windsorized to 2.5 standard deviations above the mean ($n = 14$).

2.2.2.2. Calculation of cortisol features. Calculated cortisol features are depicted in Fig. 1 and described below (formulas are provided in Table 3). Throughout formulas, *T* represents cortisol concentration estimates at each collection time and *t* represents the time (e.g., 7:30 a.m.) of sample collection.

2.2.2.2.1. Total area under the curve (AUC). Area under the curve was calculated with respect to ground (AUC*g*) and increase (AUC*i*; [23]. AUCg covers the entire area under the curve and was calculated as follows:
$T1 + T2/2 \times (t2 - t1) + T2 + T3/2 \times (t3 - t1) + \ldots\ T5 + T6/2 \times (t6 - t5)$ using all time points (SPAN T1-T6; MLSRA T1-T5).

AUC*i* only includes the increase from waking cortisol (T1) and was calculating by subtracting total time multiplied by the concentration at T1 (e.g., in SPAN: $(t6 - t1) \times T1$) from the value for AUCg. For both AUCg and AUC*i* calculations, if a sample estimate for T3-T6 was not available (e.g., not provided by the participant), it was replaced with the previous time and value (e.g. if T4 was previously excluded, T3 and t3 were used instead).[2] A day was excluded if T1 or T2 was not included in the raw data or if the day had fewer than 3 included time points. The SPAN and MLSRA datasets had final analytic samples of 131 and 90 participants, respectively, for both AUCg and AUC*i*.

2.2.2.2.2. Cortisol awakening response (CAR). Cortisol awakening response (CAR) was measured according to concentration difference ($\text{CAR}_{\text{Incline}}$) and area under the curve (CAR_{AUC}). $\text{CAR}_{\text{Incline}}$ was calculated by subtracting T1 from T2 ($T2 - T1$). CAR_{AUC} was calculated as follows: $(T2 + T1/2) \times (t2 - t1)$. For both measures, a day was excluded if its values for either T1, T2, t1, or t2 were excluded from the raw data (as described above). This resulted in a final analytic sample of 131 participants for both $\text{CAR}_{\text{Incline}}$ and CAR_{AUC} in the SPAN dataset. In the MLSRA dataset, there was a final analytic sample of 83 participants for $\text{CAR}_{\text{Incline}}$ and CAR_{AUC}.

2.2.2.2.3. Early Decline (ED) slope. Early Decline (ED) represents the steepest decline in cortisol levels, from peak to T3. ED_{Slope} was calculated using the following formula: $T3 - T2/_{t3\,-\,t2}$. If T1 was higher than T2, T1 and t1 were used in the calculation instead. Exclusion criteria were the same as CAR. Additionally, a day was excluded if T3 was excluded or if t3 was not recorded. There was a final analytic sample of 127 participants for ED_{Slope} in the SPAN datasets. ED_{Slope} was not calculated in the MLSRA dataset due to the short time period between sample collection for T2 (i.e., 30 min after waking) and T3 (i.e., 1 h post waking).

2.2.2.2.4. Late decline (LD) slope and late decline area under the curve. Late Decline (LD) represents the slope of the decline in cortisol during the later hours of the day where cortisol is declining less steeply from T3 to the end of the night (T_{final}, T6 in SPAN, T5 in MLSRA; [24,25]. This was calculated with the equation $T_{final} - T3/_{t_{final}\,-\,t3}$. Area under the curve for late decline was also calculated (LD_{AUC}) using the equation $T3 + T4/2 \times (t4 - t3) + T4 + T5/2 \times (t5 - t4) + T5 + T6/2 \times (t6 - t5)$. For both measures, if either T3 or t3 were excluded, T4 and t4 were used instead. Similarly, if the last time point (T_{final} or t_{final}) was excluded (T6 in SPAN, T5 in MLSRA), the previous concentration and time were used instead. If both T3 or t3, and T_{final} or t_{final} were missing, the day was excluded. There was a final analytic sample of 137 participants for LD_{Slope} and LD_{AUC} for the SPAN dataset. There was a final analytic sample of 90 participants for LD_{Slope} and LD_{AUC} for the MLSRA dataset.

2.2.2.2.5. Maximum decline. Maximum Decline represents the decline in concentration from the highest to the lowest value. Because of the cortisol awakening response, T2 is expected to be the peak value, therefore the equation *T2* – [minimum concentration] was used. However, to account for possible delays in sample collection upon waking, if T1 was higher than T2, T1 was used for the calculation. Exclusion criteria were the same as reported for CAR variables. Additionally, a day was excluded if there were 3 or fewer included samples for that day. For Maximum Decline there was a final analytic sample of 129 participants for the SPAN dataset and 90 participants for the MLSRA dataset.

2.2.2.2.6. Diurnal slope. Diurnal Slope was calculated as the slope between cortisol concentration at T2 and the final collection time point using the equation $T_{final} - T2/_{T_{final}\,-\,t2}$. As with the Maximum Decline calculations, T2 and t2 were replaced with T1 and t1 if T1 was a higher value. Because cortisol is relatively stable in the evening, if T6/t6 was missing or excluded, T5 was used in the calculation instead. Criteria for exclusion were the same as CAR variables. In addition to these, if both T5/t5 and T6/t6 were missing or excluded the day was excluded. There was a final analytic sample of 129 participants for Diurnal Slope for the SPAN dataset and 90 participants for the MLSRA dataset.

2.2.2.3. SPAN protocol adherence. Protocol adherence was assessed using the compliance score system described by Ref. [24]. Raw difference scores (between self-report and MEMS® Cap times) were highly skewed at every time point; therefore, a point system was used to split participants into high, medium, and low protocol adherence categories. For each sample, the absolute value of the difference between the self-reported time and the MEMS® Cap time was calculated and given a score from 0 to 3. Differences of less than 5 min were given a score of 3, differences of 5–10 min were given a score of 2, differences of 10–15 min were given a score of 1, and differences of greater than 15 min were given a score of 0. Thus, higher scores indicated greater protocol adherence. Scores were averaged across all available time points, and participants were split into the three adherence categories based on the tertiles of the distribution (see Supplemental Table 5).

2.2.3. Statistical analyses

2.2.3.1. Reliability of cortisol features. Participants were required to have each cortisol index measured on at least 2 days to be included in analyses for that feature. All data analysis was conducted in R using the packages rptR, lme4, and lmerTest [26–28]. For each of the multi-timepoint cortisol features as well as individual timepoints (SPAN $n = 6$, MLSRA $n = 5$), intraclass correlation coefficients (ICCs) across the 2–3 days were estimated using linear mixed models. Days were entered as a fixed effect and persons were entered as a random effect to account for within-subject correlation. All models were first estimated using restricted maximum likelihood (REML) without any covariates. ICCs were calculated as the ratio of between-subject variance (${\sigma_p}^2$) to the total variance (${\sigma_p}^2 + {\sigma_d}^2$), or:

[2] This was determined after establishing that correlations across these time points are highly correlated among individuals with complete data.

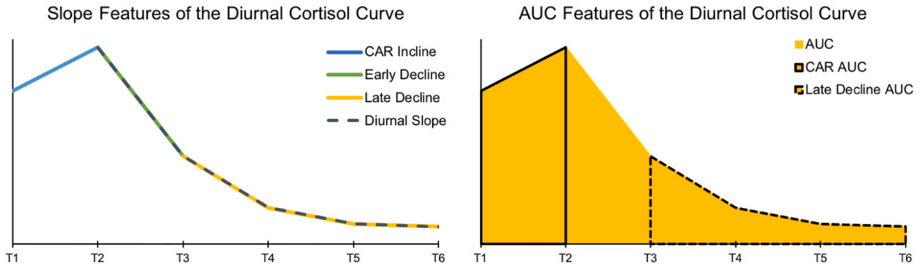


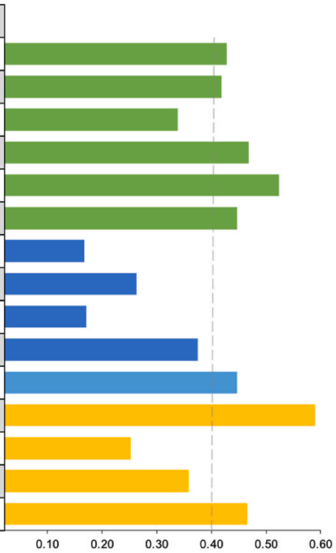
Fig. 1. Graphical depiction of features of the cortisol curve calculated in this study. Not shown is the Maximum Decline feature, which was calculated as the lowest measured cortisol value (usually, T6) from the largest measured cortisol value (usually, T2). CAR = Cortisol awakening response, AUC = Area under the curve.

Table 3

a. ICC values for cortisol features evaluated in the SPAN dataset (a) and MLSRA dataset (b). Bolded values are >0.40, indicating fair reliability. Throughout formulas, *T* represents cortisol concentration estimates at each collection time and *t* represents the time (e.g., 7:30 a.m.) of sample collection.

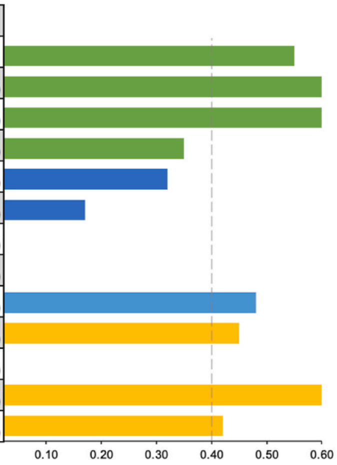
(a)

Variable Name	Formula	ICC (95% CI)
T1 (Wakeup)	T1	**0.43** (0.32 - 0.53)
T2 (Peak)	T2	**0.42** (0.31 - 0.54)
T3	T3	0.34 (0.22 - 0.45)
T4	T4	**0.47** (0.37 - 0.57)
T5	T5	**0.52** (0.42 - 0.61)
T6 (Bedtime)	T6	**0.45** (0.34 - 0.55)
CAR Incline	$T2 - T1$	0.17 (0.03 - 0.30)
Early Decline (ED) Slope	$^{T3 - T2}/_{t3 - t2}$	0.26 (0.14 - 0.40)
Late Decline (LD) Slope	$^{T6 - T3}/_{t6 - t3}$	0.17 (0.06 - 0.28)
Diurnal Slope	$^{T6 - T2}/_{t6 - t2}$	0.38 (0.24 - 0.49)
Maximum Decline (MD)	$T2 - [minimum\ concentration]$	**0.45** (0.32 - 0.58)
AUC*g*	$^{T1+T2}/_{2} \times (t2 - t1) + {}^{T2+T3}/_{2} \times (t3 - t2) + \ldots {}^{T5+T6}/_{2} \times (t6 - t5)$	**0.59** (0.48 - 0.55)
AUC*i*	$AUCg - (t6 - t1) \times T1$	0.25 (0.12 - 0.38)
CAR AUC	$^{T2+T1}/_{2} \times (t2 - t1)$	0.36 (0.23 - 0.48)
LD AUC	$^{T3+T4}/_{2} \times (t4 - t3) + {}^{T4+T5}/_{2} \times (t5 - t4) + {}^{T5+T6}/_{2} \times (t6 - t5)$	**0.47** (0.36 - 0.57)



(b)

Variable Name	Formula	ICC (95% CI)
T1 (Wakeup)	T1	**0.55** (0.39 - 0.68)
T2 (Peak)	T2	**0.65** (0.51 - 0.75)
T3	T3	**0.75** (0.63 - 0.82)
T4	T4	0.35 (0.16 - 0.54)
T5 (Bedtime)	T5	0.32 (0.12 - 0.51)
CAR Incline	$T2 - T1$	0.17 (0.00 - 0.39)
Late Decline (LD) Slope	$^{T5 - T3}/_{t5 - t3}$	0.00 (0.00 - 0.21)
Diurnal Slope	$^{T5 - T2}/_{t5 - t2}$	0.00 (0.00 - 0.21)
Maximum Decline (MD)	$T2 - [minimum\ concentration]$	**0.48** (0.31 - 0.63)
AUC*g*	$^{T1+T2}/_{2} \times (t2 - t1) + {}^{T2+T3}/_{2} \times (t3 - t2) + \ldots {}^{T4+T5}/_{2} \times (t5 - t4)$	**0.45** (0.26 - 0.60)
AUC*i*	$AUCg - (t5 - t1) \times T1$	0.00 (0.00 - 0.21)
CAR AUC	$^{T2+T1}/_{2} \times (t2 - t1)$	**0.63** (0.47 - 0.75)
LD AUC	$^{T3+T4}/_{2} \times (t4 - t3) + {}^{T4+T5}/_{2} \times (t5 - t4)$	**0.42** (0.22 - 0.57)



$$ICC = \frac{{\sigma_p}^2}{{\sigma_p}^2 + {\sigma_d}^2}$$

where is the between-day variance of the cortisol feature.

ICCs can range from 0 to 1. As a larger ratio indicates a greater proportion of the total variance is attributable to individual variability (or variability due to an effect of persons), cortisol features with larger ICCs may capture more trait-like effects. Consistent with recommended standards for reliability in psychology, we interpret reliability as poor (ICC<0.40), fair (0.40–0.60), good (0.60–0.75), or excellent (0.75; [29].

2.2.3.2. Post-hoc analyses in SPAN. Three post-hoc analyses were conducted in the SPAN dataset. *First,* we examined whether saliva sample protocol adherence, common demographic factors (i.e., self-reported sex and race), and exposure to childhood maltreatment may influence reliability. To this end, we estimated reliability when including recruitment features (i.e., sex, race, CTQ category [high, low]) as well as protocol adherence (low, medium, or high) as fixed effect covariates. 95% confidence intervals of the ICC value were computed before and after the addition of each covariate and were compared to assess whether these factors influenced reliability estimates. In addition, ICCs were generated within separate groups across these variables (e.g., low, medium, high protocol adherence).

Second, as point estimates of cortisol values could plausibly be influenced by the precision of estimated concentrations, we binned data into decile (i.e., 0–10th percentile, 11th – 20th percentile, etc.) and quintile (i.e., 0–20th percentile, 21–40th percentile, etc.) groups and recalculated ICCs to evaluate whether data binning (i.e., smoothing) improved diurnal cortisol reliability estimates.

Third and finally, AUC*g* showed the highest reliability in meta-analyses and across our studies (SPAN and MLSRA; see **Results** below). As AUC*g* uses all collected data, it places the highest burden on participants and researchers. Therefore, we examined whether subsets of timepoints can be used to generate a reliable approximation of diurnal AUC*g*, i.e., a measure that is both strongly correlated with the full AUC*g* estimation and is also reliable. To this end, using only data from days when all 6 times points were available (n = 254 days), we estimated the reliability of AUC*g* when systematically removing time points (e.g., T6, T6 + T5, etc.) and whether AUC*g* estimates derived from these iterations were correlated with AUC*g* estimated using all samples.

3. Results

3.1. Meta-analytic reliability of cortisol features

Ten articles describing 11 studies were identified for inclusion in the meta-analysis, yielding 48 ICC estimates for 15 cortisol features derived from 3307 unique participants (Table 1; [24,25,30–37]. Five cortisol features (i.e., Wakeup/T1, Peak/T2, Bedtime/T6, $\text{CAR}_{\text{Incline}}$, and AUC*g*) were reported in 4 or more studies and were subjected to meta-analysis. Meta-analytic ICCs ranged from poor ($\text{CAR}_{\text{Incline}}$: 0.29, 95% C.I.: [0.19, 0.37]) to good (AUC*g*: 0.66, [0.59, 0.71]; Fig. 2).

3.2. Cortisol reliability in SPAN and MLSRA

Multivariate Diurnal Features. As displayed in Table 3, ***AUCg***, ***Max Decline***, and ***Late Decline***$_{\boldsymbol{AUC}}$ all had fair reliability (ICCs: 0.42–0.59) in both the SPAN and MLSRA samples. Of these metrics, AUC*g* displayed the highest reliability across both studies. ***CAR***$_{\boldsymbol{AUC}}$ had good reliability in MLSRA (ICC = 0.63), but poor reliability in SPAN (ICC = 0.36). All other cortisol features (i.e., ***AUCi, CAR***$_{\textbf{Incline}}$***, ED***$_{\textbf{Slope}}$***, LD***$_{\textbf{Slope}}$***, Diurnal Slope***) demonstrated poor reliability across studies (ICCs = 0.00–0.38).

Univariate Diurnal Features (i.e., Individual Timepoints). Across both studies, individual timepoints had reliability values that were poor-excellent ranging from 0.32 to 0.75 (Table 3). In SPAN, all samples had fair reliability, with the exception of T3 (2.5 h after waking; ICC = 0.34); T5 (i.e., 12 h after waking) was the most reliable (ICC = 0.52). In MLSRA, T1-T3 measures (i.e., immediately upon waking, 30 min after waking, 1 h affect waking) demonstrated fair-excellent reliability, while the T4 (afternoon) and T5 (just before bed) samples showed poor reliability (ICCs <0.36).

3.2.1. Additional analyses in SPAN

3.2.1.1. Protocol adherence. Sex, but not race or CTQ score, was associated with protocol adherence (sex: c^2 = 7.014, *p*=0.03; race: c^2 = 0.473, *p*=.789; CTQ: c^2 = 2.759, *p*=.599; see Supplemental Fig. 2). Men and women did not differ in the low or high adherence groups (c^2s = 1.08, *ps > .180*), but there was a non-significant trend toward more men than women in the medium adherence group (c^2 = 7.02, *p*=0.08).

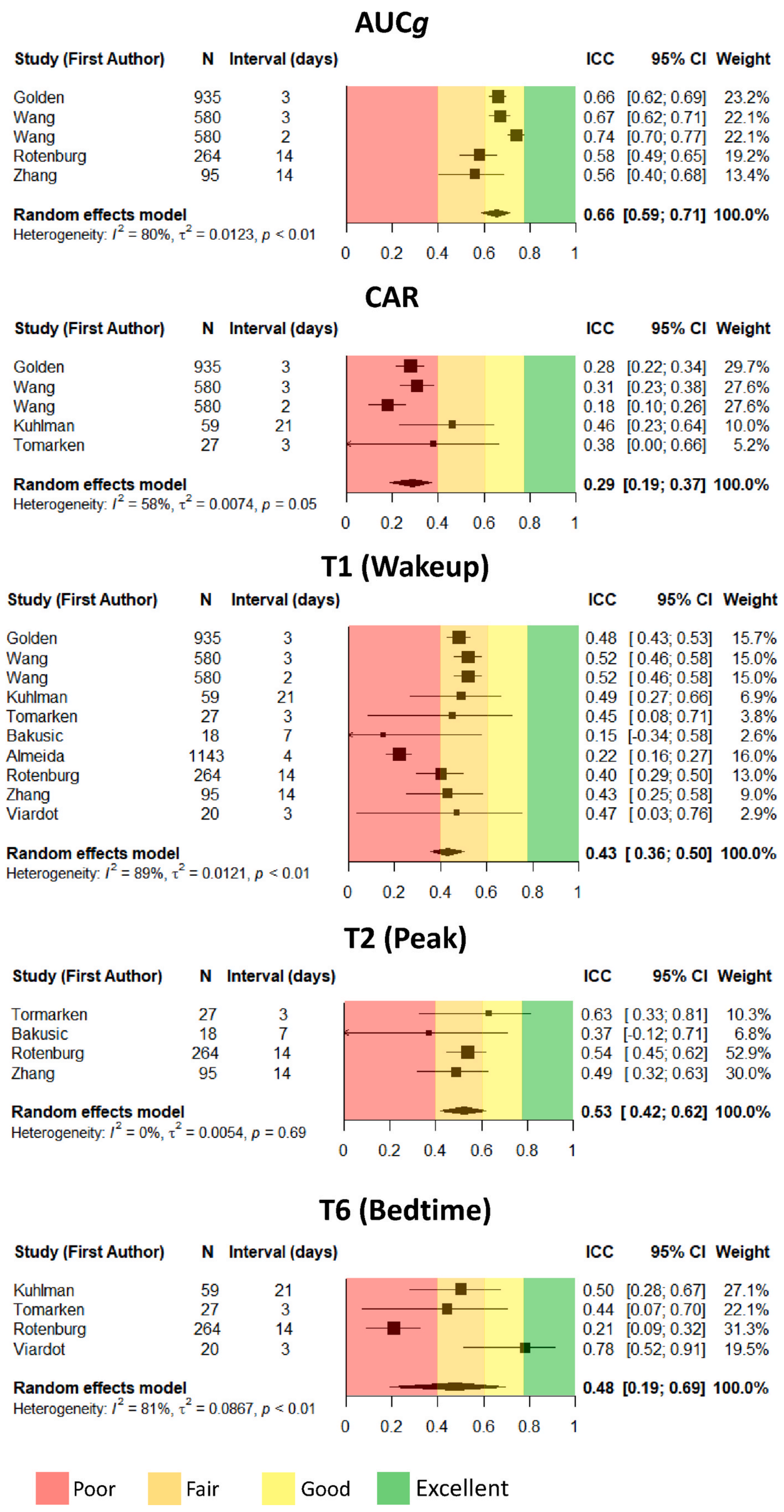
3.2.1.2. Consideration of factors that May influence reliability. The addition of the covariates sex, race, CTQ category, or protocol adherence did not improve ICC estimates generated from our linear mixed models as determined by overlapping 95% confidence intervals. Further, ICCs generated within separate groups across these variables (e.g., low, medium, high protocol adherence) yielded no evidence of differential reliability (i.e., estimates residing within 95% confidence intervals; Supplemental Table 6).

3.2.1.3. Data binning. Cortisol feature data binned into deciles (10 categories of 10%) or quintiles (5 categories of 20%) did not alter reliability estimates derived from our primary analysis for any cortisol feature (i.e., point estimates within 95% CIs; Supplemental Table 7).

3.2.1.4. Determination of minimum time samplings needed to maintain reliability. Correlations between total AUC*g* and AUC*g* calculated using a subset of data points ranged from 0.754 (AUC*g* calculated using samples T5 and T6 only) to 0.980 (AUC*g* calculated using samples T3-T6; Fig. 3a). As a reference data set, the ICC was calculated including only those days in which 6 timepoints could be included. The ICC for this full data set was 0.68 [0.57–0.76]. When AUC*g* was calculated using samples T3-T6, the ICC was 0.65 [0.552–0.741]. When T6 was excluded, the ICC was 0.56, and when both the CAR samples and T6 were excluded, the ICC was 0.50. Fig. 3b shows the ICCs for each data set.

4. Discussion

Our meta-analysis and investigation of diurnal cortisol reliability in two independent samples revealed that commonly assessed features of diurnal cortisol are highly variable in their test-retest reliability across nearby days (ICC range = 0.0–0.75). Some measures of diurnal cortisol function showed sufficient test-retest reliability for individual differences research (e.g., AUCg, Maximum Decline; individual timepoints) while others did not (e.g., $\text{CAR}_{\text{Incline}}$). Notably, however, even the most reliable diurnal cortisol features showed less than desirable reliability for trait-related individual differences research (ICCs = ~0.40–0.60). Overall, these findings suggest that the reliability of diurnal cortisol features may contribute to equivocal findings arising from studies evaluating links between diurnal cortisol and individual differences (e.g., health, stress exposure, etc.). These results encourage skepticism toward observed correlations between diurnal cortisol features and outcomes, especially in small samples. Following best practice guidelines [38], including the aggregation of data across multiple days to isolate trait-related variability has been shown to improve the reliability of some diurnal cortisol metrics, including AUGg [15]. However, the low day-to-day reliability of some commonly used (e.g., $\text{CAR}_{\text{incline}}$), likely precludes their use as trait markers, even when many days are available. Practically, more reliable features of diurnal cortisol should be prioritized in studies (e.g., AUC*g*), and larger sample sizes are needed to buttress against measurement error and non-trait-related variability of diurnal cortisol. Finally, identifying tractable factors associated with diurnal cortisol reliability (e.g., sample collection, study design, analytic) may be leveraged to improve measurement.



AUC*g*

Study (First Author)	N	Interval (days)	ICC	95% CI	Weight
Golden	935	3	0.66	[0.62; 0.69]	23.2%
Wang	580	3	0.67	[0.62; 0.71]	22.1%
Wang	580	2	0.74	[0.70; 0.77]	22.1%
Rotenburg	264	14	0.58	[0.49; 0.65]	19.2%
Zhang	95	14	0.56	[0.40; 0.68]	13.4%
Random effects model			**0.66**	**[0.59; 0.71]**	**100.0%**

Heterogeneity: $I^2 = 80\%$, $\tau^2 = 0.0123$, $p < 0.01$

CAR

Study (First Author)	N	Interval (days)	ICC	95% CI	Weight
Golden	935	3	0.28	[0.22; 0.34]	29.7%
Wang	580	3	0.31	[0.23; 0.38]	27.6%
Wang	580	2	0.18	[0.10; 0.26]	27.6%
Kuhlman	59	21	0.46	[0.23; 0.64]	10.0%
Tomarken	27	3	0.38	[0.00; 0.66]	5.2%
Random effects model			**0.29**	**[0.19; 0.37]**	**100.0%**

Heterogeneity: $I^2 = 58\%$, $\tau^2 = 0.0074$, $p = 0.05$

T1 (Wakeup)

Study (First Author)	N	Interval (days)	ICC	95% CI	Weight
Golden	935	3	0.48	[0.43; 0.53]	15.7%
Wang	580	3	0.52	[0.46; 0.58]	15.0%
Wang	580	2	0.52	[0.46; 0.58]	15.0%
Kuhlman	59	21	0.49	[0.27; 0.66]	6.9%
Tomarken	27	3	0.45	[0.08; 0.71]	3.8%
Bakusic	18	7	0.15	[-0.34; 0.58]	2.6%
Almeida	1143	4	0.22	[0.16; 0.27]	16.0%
Rotenburg	264	14	0.40	[0.29; 0.50]	13.0%
Zhang	95	14	0.43	[0.25; 0.58]	9.0%
Viardot	20	3	0.47	[0.03; 0.76]	2.9%
Random effects model			**0.43**	**[0.36; 0.50]**	**100.0%**

Heterogeneity: $I^2 = 89\%$, $\tau^2 = 0.0121$, $p < 0.01$

T2 (Peak)

Study (First Author)	N	Interval (days)	ICC	95% CI	Weight
Tormarken	27	3	0.63	[0.33; 0.81]	10.3%
Bakusic	18	7	0.37	[-0.12; 0.71]	6.8%
Rotenburg	264	14	0.54	[0.45; 0.62]	52.9%
Zhang	95	14	0.49	[0.32; 0.63]	30.0%
Random effects model			**0.53**	**[0.42; 0.62]**	**100.0%**

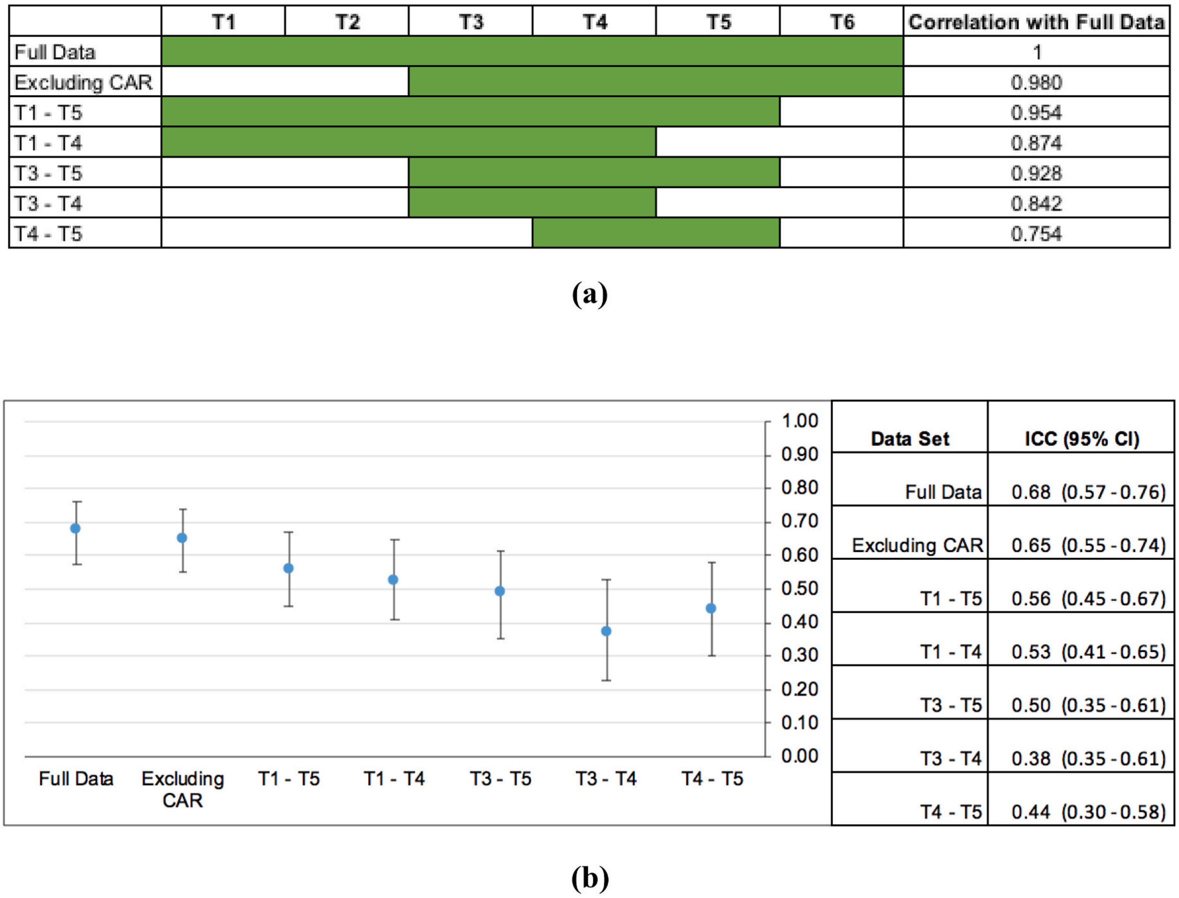
Heterogeneity: $I^2 = 0\%$, $\tau^2 = 0.0054$, $p = 0.69$

T6 (Bedtime)

Study (First Author)	N	Interval (days)	ICC	95% CI	Weight
Kuhlman	59	21	0.50	[0.28; 0.67]	27.1%
Tomarken	27	3	0.44	[0.07; 0.70]	22.1%
Rotenburg	264	14	0.21	[0.09; 0.32]	31.3%
Viardot	20	3	0.78	[0.52; 0.91]	19.5%
Random effects model			**0.48**	**[0.19; 0.69]**	**100.0%**

Heterogeneity: $I^2 = 81\%$, $\tau^2 = 0.0867$, $p < 0.01$

Fig. 2. Meta-analysis forest plot displaying the estimate of reliability for 5 cortisol features: AUC*g*, cortisol awakening response (CAR), wakeup, peak, and bedtime (last sample of the day). The first column labels each article by the first author's last name. All articles are provided in the references of this manuscript. Error bars indicate 95% confidence intervals. Larger boxes indicate studies with larger N.



	T1	T2	T3	T4	T5	T6	Correlation with Full Data
Full Data	🟩	🟩	🟩	🟩	🟩	🟩	1
Excluding CAR			🟩	🟩	🟩	🟩	0.980
T1 - T5	🟩	🟩	🟩	🟩	🟩		0.954
T1 - T4	🟩	🟩	🟩	🟩			0.874
T3 - T5			🟩	🟩	🟩		0.928
T3 - T4			🟩	🟩			0.842
T4 - T5				🟩	🟩		0.754

Data Set	ICC (95% CI)
Full Data	0.68 (0.57 - 0.76)
Excluding CAR	0.65 (0.55 - 0.74)
T1 - T5	0.56 (0.45 - 0.67)
T1 - T4	0.53 (0.41 - 0.65)
T3 - T5	0.50 (0.35 - 0.61)
T3 - T4	0.38 (0.35 - 0.61)
T4 - T5	0.44 (0.30 - 0.58)

Fig. 3. **(a)** A summary of the data sets compared to the full data set (T1 - T6) and their correlations with the full data. **(b)** ICCs of datasets with one or more time points removed. "Full Data" included only those participants for whom all 6 timepoints could be included for at least 2 days.

4.1. *AUCg: a fairly reliable index of diurnal cortisol*

Total area under the curve with respect to ground (AUC*g*), which reflects total cortisol output over the day, had the highest reliability across analyses (meta-analysis: ICC = 0.66, Fig. 2; SPAN Study ICC = 0.59; MLSRA Study ICC = 0.45, Table 3). AUC*g* has been widely studied in the context of behavior (e.g., cognitive performance), risk (e.g., environmental stress), and disease (e.g., depression). However, assessing diurnal cortisol using AUC*g* has some practical limitations: it requires multiple samples from participants throughout the day and can be more costly to analyze in densely sampled studies. For example, while there is widespread variability in the number of samples collected throughout the day and their times of collection across studies, the MacArthur Research Network on SES and Health guidelines recommended collection at 6 timepoints each day for 3–4 days (macfound.org). Such collection across the day is burdensome to participants, prone to potential error in collection time, and costly from a research perspective (staff burden, analytic, storage). Notably, our analyses in the SPAN dataset suggest that AUC*g* can be approximated (i.e., r > 0.92, Fig. 3a) reliably with as few as three afternoon samples (i.e., 2.5 h after waking [T3], 8 h after waking [T4], 12 h after waking [T5]; Fig. 3b). This may facilitate acquisition in larger studies, which are more adequately powered to confront the reliability challenges faced by diurnal cortisol metrics.

4.2. *Cortisol awakening response*

The cortisol awakening response (CAR) is a large increase in cortisol within the first hour of waking from sleep. It is typically measured by calculating a difference score (i.e., slope; here called $CAR_{Incline}$), though area under the curve approaches have also been used (here called CAR_{AUC}). CAR has been widely studied in the context of health and related risk factors, although the literature on the CAR is inconsistent and some studies show opposite effects [39]. There has been speculation that these inconsistencies may arise from protocol adherence problems (e.g., delayed sample provisions for the initial sample acquired immediately upon awakening; [40]. Here, despite sampling protocols that involved the use of timed collection devices (medication adherence caps) and evidence of good adherence, we found that the $CAR_{Incline}$ was not reliable across days in both our meta-analysis (ICC = 0.29) and independent samples (ICCs = 0.17). These data suggest that associations observed with $CAR_{Incline}$ are contaminated by poor measurement consistency and are more likely to generate false positive or negative results.

Notably, the use of an area-under-the-curve approach, CAR_{AUC}, improved the reliability of the CAR in our two independent samples to a level that approached or exceeded acceptable reliability. As such, these data suggest that the poor reliability of $CAR_{Incline}$ may reflect inconsistency introduced by difference scores as opposed to reflect large state-related variability in the CAR. More broadly, these data highlight the potential utility of CAR_{AUC} to capture the CAR, though additional studies of its reliability are needed. It remains possible that highly controlled settings that can carefully monitor collection times (e.g., overnight stays with staff support to monitor waking and collection) may improve the reliability of CAR, though this may compromise ecological validity and introduce other confounds.

4.3. *Other indices of diurnal cortisol*

Individual cortisol time points were broadly reliable in the meta-analysis as well as both independent studies. Reliability of cortisol measures in the beginning of the day, specifically T2 and T3, was marginally better in the MLSRA dataset than in the SPAN dataset. In both studies, T2 (ICC in SPAN: 0.42; ICC in MLSRA: 0.65) was sampled at 30 min after waking. The two studies sampled from very different age groups (i.e. middle age vs. old age), and this may account for the observed difference in reliability of the peak cortisol response, which may be more variable in old age [30]. The difference in ICC values for T3, however, is more likely to be explained by methodological differences – as T3 in SPAN was defined as 2.5 h post-waking, but was defined as 1 h post-waking in MLSRA.

Across the SPAN and MLSRA studies, less frequently used indices of cortisol including Maximum Decline and Late Decline AUC showed evidence of fair reliability. For the most part, aside from Maximum Decline, cortisol features calculated using difference scores (e.g., Early Decline Slope, Late Decline Slope) were characterized by poor reliability.

4.4. Modifiers of reliability

Study design issues and protocol adherence, as well as other factors (e.g., sex, race, childhood maltreatment), are thought to influence associations between cortisol and other outcomes [40,41]. We found no evidence that protocol adherence or other between participant differences (i.e., sex, race, childhood maltreatment) were associated with different reliability in the SPAN study. Nonetheless, it remains plausible that important modifiers of reliability may be identified in future work. The SPAN cortisol protocol was relatively rigorous (e.g., medication adherence caps logging time alongside participant self-report) and protocol adherence was prioritized for selecting samples for analysis (e.g., 4 days were collected and the 3 most adherent days were selected for assays); it is possible that greater deviations in participant protocol adherence may reduce reliability further and that considering such large deviations in protocol adherence may improve reliability in other studies. Finally, we found no evidence that binning participants into cortisol groups (e.g., grouping participants into deciles) improved reliability estimates relative to using estimated concentrations suggesting that smoothing these data does not improve reliability.

4.5. Limitations

Our meta-analysis and independent studies are limited by small samples that may result in imprecise reliability. Our focus on reliability in the short term (within days in our independent studies and within a month for the meta-analysis) leaves the reliability of cortisol indices separated across time and aggregated across days unclear. Notably, however, the very low reliability across days of $\mathrm{CAR}_{\mathrm{Incline}}$ would likely make it insufficiently reliable even if aggregated across time. While our 2 data collection studies engaged in many recommended best practices for cortisol collection (e.g., medication adherence caps in addition to self-reported time of collection), our studies did not include validation of reported wake up time (e.g., a secondary observer or actigraphy), which may have attenuated the reliability of cortisol awakening response.

Furthermore, The existing literature is limited in its reports of cortisol features with few studies reporting on many indices that are commonly used, which restricted our meta-analysis to 5 cortisol features. Finally, our analyses focused on the reliability of diurnal cortisol features and did not address the reliability of acute stress-induced elevations in cortisol or the reliability/stability of diurnal cortisol data aggregated across days.

4.6. Future directions

Given the large number of diurnal cortisol features that may be computed, it is important for existing and ongoing studies with multiple timepoints to prioritize features showing consistent evidence of at least fair reliability (e.g., AUCg, Maximum Decline). With that said, secondary analyses on multiple diurnal features with adequate correction for multiple testing would also be useful. Such analyses would allow for the detection of similarities and differences across measures and with relevant variables of interest; these data may be triangulated with reliability data to prioritize maximally effective approaches for the study of diurnal cortisol. Our study evaluated the reliability of cortisol metrics across nearby days; as the majority of studies aggregate data across multiple days, it will be important to further evaluate the reliability of aggregated metrics within weeks of one another and their long-term stability (e.g., across months and years).

Existing data and ongoing data collection efforts also offer opportunities to evaluate factors that may be leveraged to improve reliability of diurnal cortisol estimates including study design features (e.g., sampling protocols) as well as analytic approaches (e.g., latent multivariate models and machine learning; [42,43]. Regarding analytic approaches, researchers may wish to incorporate state-trait models, which separate within-person (environmental, “state”) from between-person (“trait) variability using hierarchical linear models [44,45]. These state-trait models aim to capture the stability of the cortisol response (trait) by parsing out the variation around the individual's average. However, these models require multiple instances of consecutive collection days (e.g. 3 consecutive collection days repeated several weeks or months apart), which may not be feasible for the majority of research studies. In these cases, prioritizing more reliable indices of cortisol output may be preferable.

In the service of future work, it will be important to recruit large samples that are capable of detecting expected small effects that are measured with only fair reliable data (0.40–0.60; [46,47], while also considering the other matrices that can be leveraged to provide broad estimates of cortisol output (e.g., hair). Cortisol estimates from hair cannot provide direct data surrounding diurnal patterns and have cultural collection considerations [48], but do have evidence of reliability and stability [49] as well as significant heritability [50].

5. Conclusions

There is widespread variability in the test-retest reliability of diurnal cortisol features. The most reliable features across analyses (i.e., AUCg) should be prioritized in individual differences research. In addition, it is important for studies of complex behavior and biology to have samples that are adequately powered to detect expected small associations between variables that are only fairly reliable.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cpnec.2023.100191.

References

[1] H. Selye, The Stress of Life, McGraw-Hill Book Company, 1956.

[2] G.M. Slavich, Social safety theory: a biologically based evolutionary perspective on life stress, health, and behavior, Annu. Rev. Clin. Psychol. 16 (1) (2020) 265–295, https://doi.org/10.1146/annurev-clinpsy-032816-045159.

[3] B.S. McEwen, What is the confusion with cortisol? Chronic Stress 3 (2019), 2470547019833647 https://doi.org/10.1177/2470547019833647.

[4] B.S. McEwen, The brain on stress: toward an integrative approach to brain, body and behavior, Perspect. Psychol. Sci. : A Journal of the Association for Psychological Science 8 (6) (2013) 673–675, https://doi.org/10.1177/1745691613506907.

[5] E.K. Adam, M.E. Quinn, R. Tavernier, M.T. McQuillan, K.A. Dahlke, K.E. Gilbert, Diurnal cortisol slopes and mental and physical health outcomes: a systematic

review and meta-analysis, Psychoneuroendocrinology 83 (2017) 25–41, https://doi.org/10.1016/j.psyneuen.2017.05.018.

[6] C.J.J.G. Janssens, F.A. Helmond, V.M. Weigant, The effect of chronic stress on plasma cortisol concentrations in cyclic female pigs depends on the time of day, Domest. Anim. Endocrinol. 12 (2) (1995) 167–177, https://doi.org/10.1016/0739-7240(94)00018-V.

[7] K. Bernard, A. Frost, C.B. Bennett, O. Lindhiem, Maltreatment and diurnal cortisol regulation: a meta-analysis, Psychoneuroendocrinology 78 (2017) 57–67, https://doi.org/10.1016/j.psyneuen.2017.01.005.

[8] K.A. Dienes, N.A. Hazel, C.L. Hammen, Cortisol secretion in depressed, and at-risk adults, Psychoneuroendocrinology 38 (6) (2013) 927–940, https://doi.org/10.1016/j.psyneuen.2012.09.019.

[9] D. Rhebergen, N.C.M. Korten, B.W.J.H. Penninx, M.L. Stek, R.C. van der Mast, R. Oude Voshaar, H.C. Comijs, Hypothalamic–pituitary–adrenal axis activity in older persons with and without a depressive disorder, Psychoneuroendocrinology 51 (2015) 341–350, https://doi.org/10.1016/j.psyneuen.2014.10.005.

[10] M.R. Bhattacharyya, G.J. Molloy, A. Steptoe, Depression is associated with flatter cortisol rhythms in patients with coronary artery disease, J. Psychosom. Res. 65 (2) (2008) 107–113, https://doi.org/10.1016/j.jpsychores.2008.03.012.

[11] M.T. Tu, M.-V. Zunzunegui, R. Guerra, B. Alvarado, J.M. Guralnik, Cortisol profile and depressive symptoms in older adults residing in Brazil and in Canada, Aging Clin. Exp. Res. 25 (5) (2013) 527–537, https://doi.org/10.1007/s40520-013-0111-0.

[12] D.A. Baranger, M. Finsaas, B. Goldstein, C. Vize, D. Lynam, T. Olino, Tutorial: power analyses for interaction effects in cross-sectional regressions, PsyArXiv (2022), https://doi.org/10.31234/osf.io/5ptd7.

[13] D. Zimmerman, B. Zumbo, Resolving the issue of how reliability is related to statistical power: adhering to mathematical definitions, J. Mod. Appl. Stat. Methods: JMASM 14 (2015) 9–26, https://doi.org/10.22237/jmasm/1446350640.

[14] C. Kirschbaum, R. Steyer, M. Eid, U. Patalla, P. Schwenkmezger, D.H. Hellhammer, Cortisol and behavior: 2. Application of a Latent state-trait model to salivary cortisol, Psychoneuroendocrinology 15 (4) (1990) 297–307, https://doi.org/10.1016/0306-4530(90)90080-S.

[15] S.C. Segerstrom, I.A. Boggero, G.T. Smith, S.E. Sephton, Variability and reliability of diurnal cortisol in younger and older adults: implications for design decisions, Psychoneuroendocrinology 49 (2014) 299–309, https://doi.org/10.1016/j.psyneuen.2014.07.022.

[16] J.F. Martínez, J. Schweig, P. Goldschmidt, Approaches for combining multiple measures of teacher performance: reliability, validity, and implications for evaluation policy, Educ. Eval. Pol. Anal. 38 (4) (2016) 738–756, https://doi.org/10.3102/0162373716666166.

[17] L.J. Cronbach, L. Furby, How we should measure "change": or should we? Psychol. Bull. 74 (1970) 68–80, https://doi.org/10.1037/h0029382.

[18] M. Harrer, P. Cuijpers, T.A. Furukawa, D.D. Ebert, Doing Meta-Analysis with R: A Hands-On Guide, first ed., Chapmann & Hall/CRC Press, 2021. https://ebin.pub/doing-meta-analysis-with-r-a-hands-on-guide-1nbsped-0367610078-9780367610074-e-4393992.html.

[19] T.F. Oltmanns, M.M. Rodrigues, Y. Weinstein, M.E.J. Gleason, Prevalence of personality disorders at midlife in a community sample: disorders and symptoms reflected in interview, self, and informant reports, J. Psychopathol. Behav. Assess. 36 (2) (2014) 177–188, https://doi.org/10.1007/s10862-013-9389-7.

[20] D.P. Bernstein, L. Fink, L. Handelsman, J. Foote, Childhood Trauma questionnaire. https://doi.org/10.1037/t02080-000, 1998.

[21] L.A. Sroufe, B. Egeland, E.A. Carlson, W.A. Collins, The Development of the Person: the Minnesota Study of Risk and Adaptation from Birth to Adulthood, Guilford Publications, 2005, p. 384, xvi.

[22] E.S. Young, A.K. Farrell, E.A. Carlson, M.M. Englund, G.E. Miller, M.R. Gunnar, G. I. Roisman, J.A. Simpson, The dual impact of early and concurrent life stress on adults' diurnal cortisol patterns: a prospective study, Psychol. Sci. 30 (5) (2019) 739–747, https://doi.org/10.1177/0956797619833664.

[23] J.C. Pruessner, C. Kirschbaum, G. Meinlschmid, D.H. Hellhammer, Two formulas for computation of the area under the curve represent measures of total hormone concentration versus time-dependent change, Psychoneuroendocrinology 28 (7) (2003) 916–931, https://doi.org/10.1016/s0306-4530(02)00108-7.

[24] S.H. Golden, B.N. Sánchez, A.S. DeSantis, M. Wu, C. Castro, T.E. Seeman, S. Tadros, S. Shrager, A.V. Diez Roux, Salivary cortisol protocol adherence and reliability by sociodemographic features: the multi-ethnic study of atherosclerosis, Psychoneuroendocrinology 43 (2014) 30–40, https://doi.org/10.1016/j.psyneuen.2014.01.025.

[25] X. Wang, B.N. Sánchez, S.H. Golden, S. Shrager, C. Kirschbaum, A.S. Karlamangla, T.E. Seeman, A.V.D. Roux, Stability and predictors of change in salivary cortisol measures over six years: MESA, Psychoneuroendocrinology 49 (2014) 310–320, https://doi.org/10.1016/j.psyneuen.2014.07.024.

[26] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, J. Stat. Software 67 (2015) 1–48, https://doi.org/10.18637/jss.v067.i01.

[27] A. Kuznetsova, P.B. Brockhoff, R.H.B. Christensen, lmerTest package: tests in linear mixed effects models, J. Stat. Software 82 (2017) 1–26, https://doi.org/10.18637/jss.v082.i13.

[28] M.A. Stoffel, S. Nakagawa, H. Schielzeth, rptR: repeatability estimation and variance decomposition by generalized linear mixed-effects models, Methods Ecol. Evol. 8 (11) (2017) 1639–1644, https://doi.org/10.1111/2041-210X.12797.

[29] D.V. Cicchetti, S.A. Sparrow, Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior, Am. J. Ment. Defic. 86 (2) (1981) 127–137.

[30] D.M. Almeida, J.R. Piazza, R.S. Stawski, Interindividual differences and intraindividual variability in the cortisol awakening response: an examination of age and gender, Psychol. Aging 24 (4) (2009) 819–827, https://doi.org/10.1037/a0017910.

[31] J. Bakusic, S. De Nys, M. Creta, L. Godderis, R.C. Duca, Study of temporal variability of salivary cortisol and cortisone by LC-MS/MS using a new atmospheric pressure ionization source, Sci. Rep. 9 (1) (2019), 19313, https://doi.org/10.1038/s41598-019-55571-3.

[32] J. Hellhammer, E. Fries, O.W. Schweisthal, W. Schlotz, A.A. Stone, D. Hagemann, Several daily measurements are necessary to reliably assess the cortisol rise after awakening: state- and trait components, Psychoneuroendocrinology 32 (1) (2007) 80–86, https://doi.org/10.1016/j.psyneuen.2006.10.005.

[33] K.R. Kuhlman, T.F. Robles, L. Dickenson, B. Reynolds, R.L. Repetti, Stability of diurnal cortisol measures across days, weeks, and years across middle childhood and early adolescence: exploring the role of age, pubertal development, and sex, Psychoneuroendocrinology 100 (2019) 67–74, https://doi.org/10.1016/j.psyneuen.2018.09.033.

[34] S. Rotenberg, J.J. McGrath, M.-H. Roy-Gagnon, M.T. Tu, Stability of the diurnal cortisol profile in children and adolescents, Psychoneuroendocrinology 37 (12) (2012) 1981–1989, https://doi.org/10.1016/j.psyneuen.2012.04.014.

[35] A.J. Tomarken, G.T. Han, B.A. Corbett, Temporal patterns, heterogeneity, and stability of diurnal cortisol rhythms in children with autism spectrum disorder, Psychoneuroendocrinology 62 (2015) 217–226, https://doi.org/10.1016/j.psyneuen.2015.08.016.

[36] A. Viardot, P. Huber, J.J. Puder, H. Zulewski, U. Keller, B. Müller, Reproducibility of nighttime salivary cortisol and its use in the diagnosis of hypercortisolism compared with urinary free cortisol and overnight dexamethasone suppression test, J. Clin. Endocrinol. Metabol. 90 (10) (2005) 5730–5736, https://doi.org/10.1210/jc.2004-2264.

[37] Q. Zhang, Z. Chen, S. Chen, Y. Xu, H. Deng, Intraindividual stability of cortisol and cortisone and the ratio of cortisol to cortisone in saliva, urine and hair, Steroids 118 (2017) 61–67, https://doi.org/10.1016/j.steroids.2016.12.008.

[38] T. Stalder, C. Kirschbaum, B.M. Kudielka, E.K. Adam, J.C. Pruessner, S. Wüst, S. Dockray, N. Smyth, P. Evans, D.H. Hellhammer, R. Miller, M.A. Wetherell, S. J. Lupien, A. Clow, Assessment of the cortisol awakening response: expert consensus guidelines, Psychoneuroendocrinology 63 (2016) 414–432, https://doi.org/10.1016/j.psyneuen.2015.10.010.

[39] I.A. Boggero, C.E. Hostinar, E.A. Haak, M.L.M. Murphy, S.C. Segerstrom, Psychosocial functioning and the cortisol awakening response: meta-analysis, P-curve analysis, and evaluation of the evidential value in existing studies, Biol. Psychol. 129 (2017) 207–230, https://doi.org/10.1016/j.biopsycho.2017.08.058.

[40] L. Thorn, F. Hucklebridge, P. Evans, A. Clow, Suspected non-adherence and weekend versus week day differences in the awakening cortisol response, Psychoneuroendocrinology 31 (8) (2006) 1009–1018, https://doi.org/10.1016/j.psyneuen.2006.05.012.

[41] C.T. Halpern, E.A. Whitsel, B. Wagner, K.M. Harris, Challenges of measuring diurnal cortisol concentrations in a large population-based field study, Psychoneuroendocrinology 37 (4) (2012) 499–508, https://doi.org/10.1016/j.psyneuen.2011.07.019.

[42] J.E. Khoury, A. Gonzalez, R.D. Levitan, J.C. Pruessner, K. Chopra, V.S. Basile, M. Masellis, A. Goodwill, L. Atkinson, Summary cortisol reactivity indicators: interrelations and meaning, Neurobiology of Stress 2 (2015) 34–43, https://doi.org/10.1016/j.ynstr.2015.04.002.

[43] K. Yoo, M.D. Rosenberg, S. Noble, D. Scheinost, R.T. Constable, M.M. Chun, Multivariate approaches improve the reliability and validity of functional connectivity and prediction of individual behaviors, Neuroimage 197 (2019) 212–223, https://doi.org/10.1016/j.neuroimage.2019.04.060.

[44] E.A. Shirtcliff, M.J. Essex, Concurrent and longitudinal associations of basal and diurnal cortisol with mental health symptoms in early adolescence, Dev. Psychobiol. 50 (7) (2008) 690–703, https://doi.org/10.1002/dev.20336.

[45] S.F. Thompson, M. Zalewski, C.J. Kiff, L.J. Lengua, A state-trait model of cortisol in early childhood: contextual and parental predictors of stable and time-varying effects, Horm. Behav. 98 (2018) 198–209, https://doi.org/10.1016/j.yhbeh.2017.12.009.

[46] J. Brunner, P.C. Austin, Inflation of Type I error rate in multiple regression when independent variables are measured with error, Can. J. Stat. 37 (1) (2009) 33–46, https://doi.org/10.1002/cjs.10004.

[47] G.J. Matheson, We need to talk about reliability: making better use of test-retest studies for study design and interpretation, PeerJ 7 (2019), e6918, https://doi.org/10.7717/peerj.6918.

[48] L. Manns-James, A. Neal-Barnett, Development of a culturally informed protocol for hair cortisol sampling in Black women, Publ. Health Nurs. 36 (6) (2019) 872–879, https://doi.org/10.1111/phn.12668.

[49] T. Stalder, S. Steudte, R. Miller, N. Skoluda, L. Dettenborn, C. Kirschbaum, Intraindividual stability of hair cortisol concentrations, Psychoneuroendocrinology 37 (5) (2012) 602–610, https://doi.org/10.1016/j.psyneuen.2011.08.007.

[50] L. Rietschel, F. Streit, G. Zhu, K. McAloney, J. Frank, B. Couvy-Duchesne, S.H. Witt, T.M. Binz, J. McGrath, I.B. Hickie, N.K. Hansell, M.J. Wright, N.A. Gillespie, A. J. Forstner, T.G. Schulze, S. Wüst, M.M. Nöthen, M.R. Baumgartner, B.R. Walker, M. Rietschel, Hair cortisol in twins: heritability and genetic overlap with psychological variables and stress-system genes, Sci. Rep. 7 (1) (2017), https://doi.org/10.1038/s41598-017-11852-3. Article 1.