BMC Sports Science, Medicine and Rehabilitation

RESEARCH

Open Access

Test-retest reliability and validity of vagally-mediated heart rate variability to monitor internal training load in older adults: a within-subjects (repeated-measures) randomized study

Patrick Manser[1*] and Eling D. de Bruin[1,2,3]

Abstract

Background Vagally-mediated heart rate variability (vm-HRV) shows promise as a biomarker of internal training load (ITL) during exergame-based training or motor-cognitive training in general. This study evaluated the test-retest reliability of vm-HRV during exergaming in healthy older adults (HOA) and its validity to monitor ITL.

Methods A within-subjects (repeated-measures) randomized study was conducted that included baseline assessments and 4 measurement sessions. Participants played 5 exergames at 3 standardized levels of external task demands (i.e., "easy", "challenging", and "excessive") in random order for 90 s. Test-retest reliability was assessed on the basis of repeated-measures analyses of variance (ANOVA), intraclass correlation coefficients ($ICC_{3,1}$), standard errors of measurement (SEM), and smallest detectable differences (SDD). Validity was determined by examining the effect of game level on vm-HRV in the ANOVA.

Results Fourty-three HOA (67.0 ± 7.0 years; 58.1% females (25 females, 18 males); body mass index = 23.7 ± 3.0 $kg \cdot m^{-2}$) were included. Mean R-R time intervals (mRR) and parasympathetic nervous system tone index (PNS-Index) exhibited mostly good to excellent relative test-retest reliability with no systematic error. Mean SEM% and SDD% were 36.4% and 100.7% for mRR, and 44.6% and 123.7% for PNS-Index, respectively. Significant differences in mRR and PNS-Index were observed between standardized levels of external task demands, with mostly large effect sizes (mean $r = 0.847$). These results persisted irrespective of the type of neurocognitive domain trained and when only motoric and cognitive demands were manipulated while physical intensity was kept constant. The remaining vm-HRV parameters showed inconsistent or poor reliability and validity.

Conclusion Only mRR and PNS-Index demonstrated reliable measurement and served as valid biomarkers for ITL during exergaming at a group level. Nonetheless, the presence of large SEMs hampers the detection of individual changes over time and suggests insufficient precision of these measurements at the individual level. Future research should further investigate the reliability and validity of vm-HRV with a specific focus on comparing different measurement methodologies and exercise conditions, particularly focusing on ultra-short-term HRV measurements,

*Correspondence:
Patrick Manser
patrick.manser@hest.ethz.ch
Full list of author information is available at the end of the article

and investigate the potential implications (i.e., superiority to other markers of ITL or monitoring strategies?) of using vm-HRV as a biomarker of ITL.

Keywords Autonomic nervous system, Biomarkers, Exercise, Exergaming, Neurosciences

Introduction

Background

The growing need to identify and implement effective measures for the prevention of neurocognitive disorders [1] has led to the development of new approaches. While motor-cognitive training is recommended for the prevention of neurocognitive disorders by a collaborative international guideline [2], the utilization of technology to facilitate the implementation of motor-cognitive training, for instance through exergaming [3], is becoming increasingly popular. Exergames offer several advantages over conventional motor-cognitive training which promote their effectiveness and are, thus, currently considered a more promising training approach than conventional physical and/or cognitive training [4–6]. One of the key advantages of exergames is the ease of use of additional opportunities for individualized training through real-time adaptivity of task demands according to monitored parameters such as performance, measures of brain activity, or internal training load [7–9]. Nevertheless, recent systematic reviews have identified a paucity of systematic reporting and control of physical and, in particular, neurocognitive demands during exergaming in the majority of studies [10, 11]. Consequently, although training load monitoring has generally improved significantly over the past decades [12], multiple research groups have advocated that future research endeavors should prioritize the identification of reliable and valid parameters to monitor training load in order to enhance the effectiveness of physical training in general and exergame-based training in particular [10, 13–16].

The use of specific markers of "internal training load" (ITL) has been recommended to tailor exercises to the individual's capabilities and performance [10, 16, 17]. ITL during exergaming is mainly influenced by neurocognitive task demands and the physical exercise intensity [18]. In light of this, it has been recommended that physical exercise intensity be prescribed and objectively monitored [10, 11], for example in accordance with the guidelines provided by the American College of Sports Medicine [10, 19]. Assessing and monitoring neurocognitive task demands is more complex, and there is a plethora of available measurement methods [13, 20]. For exergame-based training, the use of validated game metrics or subjective self-report measures has been recommended to prescribe and monitor the motoric and cognitive demands [10]. While self-report measures have been demonstrated to generally have high levels of validity in measuring cognitive load [21], they are prone to bias [13] and a single response to assess cognitive load after completing a training session or task may be insufficient, because neurocognitive task demands may change over time [13, 20, 22]. It is therefore of great importance to identify specific and time-sensitive physiological markers for ITL, as these temporal changes should be captured by a valid marker for ITL [17].

Vagally-mediated heart rate variability (vm-HRV) has been identified as a promising parameter for monitoring ITL during simultaneous motor-cognitive training, such as exergaming [13, 23]. The "neurovisceral integration" model [24] and its advancements [25, 26] posit that vm-HRV indexes the functional integrity of the central autonomic network (CAN). The CAN regulates physiological, emotional, and cognitive responses to environmental challenges [26], which is precisely what should be reflected by an optimal marker for ITL [17, 27]. Phasic vm-HRV responses have been shown to be moderated by physical and cognitive capabilities and exercise demands [23], are sensitive to neurocognitive demands related to cognitive and mental effort in older adults [13, 28–31], and have been found suitable in distinguishing between varying intensities and durations of physical exercise [32–34].

Nevertheless, further research is necessary to ascertain the suitability of monitoring phasic vm-HRV responses as a biomarker of ITL during exergaming [23]. In addition, the validity, reliability, sensitivity to change, and applicability of vm-HRV monitoring should be evaluated. This includes determining whether vm-HRV can capture ITL changes within reasonable timeframes for adapting task demands in real-time. However, there is currently insufficient evidence regarding the validity and reliability of vm-HRV measurements during physical exercise [35, 36] or simultaneous motor-cognitive training [23] as well as ultra-short term (< 5 min) recordings [37] in HOA. To ensure that observed changes in the variable of interest are attributable to real changes rather than measurement error, it is a prerequisite to assess the test-retest reliability before further exploring the validity of vm-HRV to monitor ITL during exergame-based motor-cognitive training.

Objectives

The primary objective of this study was to evaluate the test-retest reliability of vm-HRV during exergame-based

motor-cognitive training in relation to different exergame demands in HOA. As secondary objective, the validity of vm-HRV to monitor ITL during exergame-based motor-cognitive training was investigated for parameters with acceptable test-retest reliability.

Materials and methods

Study design

A within-subjects (repeated-measures) randomized study including HOA (≥60 years of age) was conducted. The study was reported according to the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [38] (Supplementary File 1) as well as the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist for cross-sectional studies [39] (Supplementary File 2). All study procedures took place at ETH Hönggerberg (Zurich, Switzerland) and were led by investigators from our research team trained in the application of the measurement techniques and protocols. There were no changes to the methods or outcome measures after trial commencement.

After recruitment and providing written informed consent (section 'Methods - Recruitment & consent procedure'), participants were screened on eligibility (section 'Methods - Eligibility criteria' and Table 1), and the measurements were scheduled for eligible participants. At the first scheduled appointment, baseline assessments were performed, and participants were familiarized with the exergame training system "Senso" (Dividat AG, Schindellegi, Switzerland) with software version 22.4.0-360-gf9df00d5b. The system's pressure-sensitive platform uses 20 sensors to detect the position and timing of participants' movements, allowing them to control virtual exergame scenarios displayed on a frontal screen. Each participant completed 1 standardized exercise session to familiarize themselves with the exergame scenarios by play-testing each game for 2 min.

The following 4 appointments included the experimental procedures that were scheduled to take place at approximately the same time of the day (±2 h). To minimize the influence of transient confounding effects on HRV, all participants were instructed verbally and in writing to follow a normal sleep routine the day before the experiment, to avoid intense physical activities and alcohol consumption within 24 h before the measurements, and to refrain from coffee and caffeinated drinks as well as food consumption at least 2 h before the measurements [40]. No compensation was granted to the participants, but detailed feedback on individual performance as well as the study outcomes in general was provided at the end of the study.

Recruitment & consent procedure

HOA were recruited between January 2021 and June 2021 in collaboration with healthcare institutions in the larger area of Zurich by handing out leaflets to interested persons. All potential participants were fully informed about the study by trained investigators from our research team by providing verbal explanations and an information sheet. After sufficient time for considerations (i.e., at least 24 h after handing out the study information sheet, but on average around one week), interested persons willing to take part in the study provided written informed consent, were screened on eligibility in an in-person meeting, and the study sessions were scheduled.

Eligibility criteria

All eligibility criteria are detailed in Table 1.

Experimental procedures

All experimental procedures were preceded by a resting-state measurement of heart rate (HR_{rest}) and vm-HRV (section 'Measurement of HR and vm-HRV'). To account for differing exergaming conditions and distinguish between the physical and neurocognitive demands of exergaming, we evaluated the study objectives in exergaming as 'simultaneous-incorporated' (i.e., physical and cognitive demands are linked, and both change as a function of game complexity; phase 1) and 'simultaneous-additional' (phase 2) motor-cognitive training [41]. More specifically, phase 2 was based on a methodological framework for the contribution of physical and

Table 1 Description of all eligibility criteria

Inclusion criteria	Exclusion criteria
Participants fulfilling all the following inclusion criteria were eligible: • healthy (based on self-report) older adults (≥60 years) • ability to stand for at least 10 min without assistance • German speaking	The presence of any of the following criteria led to exclusion: • mobility impairments (i.e., gait, balance) that prevent from study participation • presence of neurological disorders (i.e., epilepsy, stroke, multiple sclerosis, Parkinson's disease, brain tumors, or traumatic disorders of the nervous system) • presence of any other unstable or uncontrolled diseases (e.g., uncontrolled high blood pressure and progressing or terminal cancer)

neurocognitive (i.e., game-) demands during exergaming (section 'Step 2: Development and Validation of Adaptation Loop' and figure 1 of [42]). A stepping task was used separated from game demands to keep the physical intensity constant, whereas the game demands were then manipulated. Based on this framework, changes in the overall ITL can mainly be attributed to the game demands (motoric and cognitive demands) since the physical exercise intensity is kept constant [42]. Each experimental phase included 2 measurements performed within 2 weeks.

The individual randomization of the order of games and levels of task demands was conducted by the outcome assessor, who used the randomization list from random.org. The same outcome assessor performed the test and retest measurements for individual participants. No blinding protocol was implemented. Participants were informed that different exergame demands would be applied, but were not given any information about the specific order of games and levels of task demands, nor the differences between them. Figure 1 summarizes the study procedures.

Experimental phase 1: evaluation as 'simultaneous-incorporated' motor-cognitive training

In phase 1, 5 exergames were performed in randomized order. To investigate our objectives in relation to different neurocognitive functions, 5 games that mainly demand attentional ('Simple'), executive ('Habitats', 'Targets'), and visuospatial ('Tetris') functions, as well as learning and memory ('Simon'), were played (video demonstrations of the games see [43]). Four levels of external task demand (e.g., game type, task complexity, predictability of required tasks) were applied for each game (section 'Standardized levels of external task demands').

Experimental phase 2: evaluation as 'simultaneous-additional' motor-cognitive training

First, the minimal stepping frequency to reach a moderate level of physical intensity (i.e., ranging between 40 and 59% HRR [19]) was determined using a ramp test (start level = 80 steps/min, increases of 5 steps/min every 20 s until target HR (HR_{target}) was reached). Participants followed the auditory rhythm of a metronome. Real-time HR measures (section 'Measurement of HR and vm-HRV) were averaged over each 20 s—interval. HR_{target} was calculated on basis of the Karvonen method with a target intensity of 40% HRR: HR_{target} = (maximal HR (HR_{max}) − HR_{rest}) · 0.40 + HR_{rest} [44, 45] using the age-predicted $HR_{max} = 208 - 0.7 \cdot age$.

During the block trials (Fig. 1), a variable amount of game demands was applied on top of this fixed physical intensity. The same 5 games and standardized levels of external task demands as in 'Experimental Phase 1: Evaluation as 'simultaneous-incorporated' motor-cognitive training' were used. To minimize the effect of fatigue due to physical exertion, participants rested until their HR was equal to their previous HR_{rest}. Study investigators were instructed to maintain HR_{rest} within ± 5 bpm throughout the experimental session before starting a new exergame.

Standardized levels of external task demands

For each game, adaptive task demands were first applied to minimize learning effects. Subsequently, three standardized levels of external task demands (i.e., "easy", "challenging", "excessive") were applied in randomized order for 90 s each. All external task demands were predetermined in consultation with a neuropsychologist experienced in exergame training with HOA. The aim of the 3 levels was to induce an underload in the easy condition, a challenging but feasible load in the challenging condition,

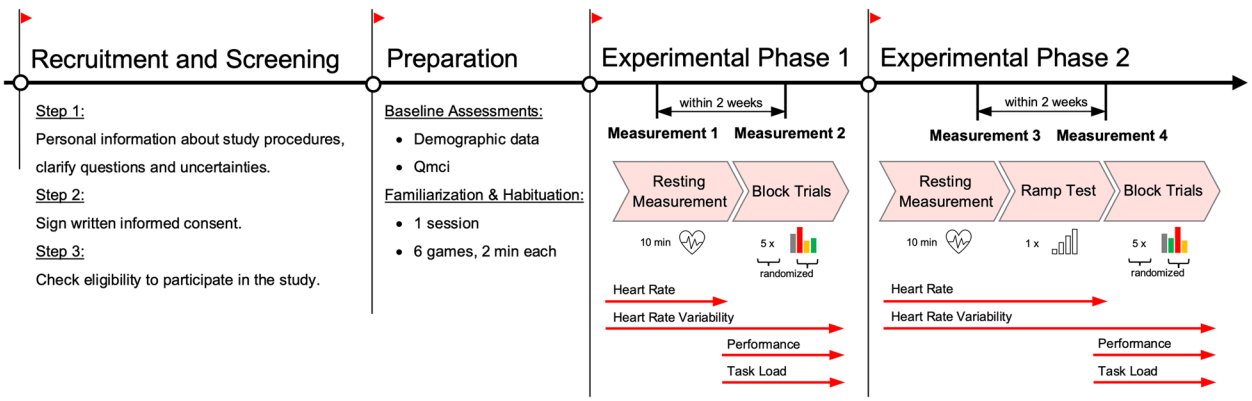

Fig. 1 Overview of the study procedures. Color coding of the block trials referring to different levels of task demands: gray = adaptive, orange = easy, green = challenging, red = excessive. Abbreviations: Qmci, Quick mild cognitive impairment screen

and an overload in the excessive condition in an average HOA. The same external loads were applied for all study participants. In the adaptive task condition, the exergame demands were adjusted using "Senso's" internal progression algorithm. This algorithm adjusts task demands in real time based on user performance, with the goal of providing optimal challenge. The game characteristics and the specific settings for all games and levels are described in Supplementary File 3.

Outcomes & data analysis

Baseline assessments

Age, sex, body mass index (i.e., body weight [kg] / (height [m])^2), physical activity behavior (i.e., measured as volume [min/week] of physical activity of at least moderate level), and self-reported intake of cardioactive medications (i.e., medications that have reported effects on HR and/or HRV; reported medications were categorized as cardioactive [yes / no] by agreement between the two authors) was assessed. Additionally, we screened the global level of cognitive functioning using the German Version [46] of the validated Quick Mild Cognitive Impairment Screen (Qmci) [47–50], which was administered and evaluated according to published guidelines [48].

Measurement of HR and vm-HRV

Resting HR and vm-HRV were measured while sitting in a comfortable position on a chair, without speaking, with both feet flat on the floor with knees at a 90° angle, hands on thighs (i.e., palms facing upward), and eyes closed [40]. Measurements were taken in a quiet room with dimmed lighting and at room temperature.

Multi-lead ECG is considered the gold standard for measuring HRV [51]. However, portable heart rate monitors (e.g., chest belts) are widely spread and have better ease of use for monitoring ITL during everyday training. Given the consistent evidence demonstrating a small amount of absolute error in HRV measurements obtained from the measurement of inter-beat-intervals through one-lead ECG via portable heart rate monitors when compared to multi-lead ECG recordings [36, 52], data was collected using a HR monitor (Polar M430) with sensor (Polar H10). The acclimatization phase was 5 min, followed by a 5 min resting measurement, the recommended duration for short-term recordings [40, 53]. The start of the resting measurement was not announced to participants [40]. In addition, R-R Intervals were continuously recorded throughout the experimental procedures.

For both, resting and on-task measurements, a sampling rate of 1000 Hz was used to provide a temporal resolution of 1 ms for each R–R interval [54]. R-R data was transmitted to Kubios HRV Premium (Kubios Oy, Kuopio, Finland, Version 3.4) for analysis. Kubios HRV is a scientifically validated software for HRV analysis and has achieved a gold-standard status in scientific research [55–58]. The automatic beat correction algorithm and noise handling provided by the software was used to correct for artifact and/or ectopic beats. The algorithm was validated for measurements at rest and was additionally tested for exercise measurements and provides reliable HRV analysis by reducing the effect of potential artefacts to a tolerable level [55]. The entire 5-min resting measurement was analyzed, while the last 60 s of on-task measurements were selected for analysis. After removing inter-beat-interval time series non-stationarities by detrending analysis using the smoothness priors method approach (settings: detrending method = smoothn priors, Lambda = 500, fc = 0.035 Hz), mean values of mainly vm-HRV indices were calculated for each segment. The mean R-R time interval (mRR), root mean square of successive RR interval differences (RMSSD), absolute power of the high-frequency (0.15–0.4 Hz; HF) band, relative power of HF (in normal units; HF [n.u.] = HF [ms^2] / (total power [ms^2] – very low frequency (0.00–0.04 Hz [ms^2])), and the Poincaré plot standard deviation perpendicular to the line of identity (SD1) were considered [37, 40, 53, 59, 60]. Additionally, the parasympathetic nervous system tone index (PNS-Index) was calculated that compares PNS activity to normal resting values [60].

Assessment of perceived task load

Participants rated their subjective task load immediately after completing each game using the NASA Task Load Index (TLX). The NASA TLX consists of 6 rating scales (subjective effort, mental demand, temporal demand, physical demand, perception of performance and frustration) ranging between '0 = very low' and '20 = very high' [61]. The raw TLX (RTLX) was calculated by summing up all sub-scores without weighting [62].

Data management

Study investigators received comprehensive training on study procedures following the Guidelines of Good Clinical Practice (GCP) and detailed working instructions. The principal investigator oversaw methodological standards and ensured quality data collection using the Castor EDC data management system (Ciwit BV, Amsterdam, The Netherlands). Data entry in electronic case report forms (eCRFs) included pre-programmed range checks. A second study investigator cross-checked all data entries before exporting the data for analysis. To minimize bias, standardized measurement procedures and participant instructions were followed according to detailed work instructions for all outcome measures.

Statistics

Statistical analyses were executed using R Version R 3.6.2 GUI 1.70 El Capitan build (7735) (© The R Foundation) in line with RStudio Version 1.2.5033 (RStudio, Inc.). First, descriptive statistics were computed for all outcome variables [63–65]. Normality distribution was checked using the Shapiro–Wilk test. The level of significance was set to $p \leq 0.05$ (2-sided). Data was reported as mean ± standard deviation for data fulfilling all the assumptions that would subsequently justify parametric statistical analyses. In case these assumptions were not met, medians (interquartile ranges) were reported. All the following statistical procedures were performed for each experimental phase separately.

Test-retest reliability of vm-HRV

For a comprehensive assessment of test-retest reliability, a 3-level approach was adopted as recommended by Weir [66]. Only data from participants with complete datasets of high-quality data (i.e., less than 5% of beats corrected by the automatic beat correction algorithm of Kubios HRV Premium) were included in the analysis.

First, a 2-way repeated-measures analyses of variance (ANOVA; timepoints of measurement X standardized levels of external task demands) or (in case of non-parametric analyses) a robust ANOVA using the nparLD package [67] was computed to examine systematic error. In case of violation of the assumption of sphericity (assessed using Mauchly's test), Greenhouse–Geisser corrections were applied [65].

Second, relative reliability was assessed by calculating intraclass correlation coefficients ($ICC_{3,1}$) estimates and their 95% confidence intervals for the agreement between repeated measurements [66, 68, 69]. ICC's were interpreted as representing poor ($ICC_{3,1} < 0.50$), fair ($0.5 \leq ICC_{3,1} < 0.75$), good ($0.75 \leq ICC_{3,1} < 0.9$), or excellent ($ICC_{3,1} \geq 0.9$) agreement [68].

Third, absolute reliability was assessed by calculating standard errors of measurement (SEM) = standard error (of both measurements) * $\sqrt{(1 - ICC_{3,1})}$; expressed as mean-normalized SEM (SEM%; SEM% = 100 * SEM / mean, combined mean value of both measurements) and the smallest detectable differences (SDD) = SEM * 1.96 *$\sqrt{2}$; expressed as mean-normalized SDD (SDD% = 100 * SDD / mean, combined mean value of both measurements) [66].

Validity of vm-HRV to monitor ITL

Only outcome measures with acceptable test-retest reliability (i.e., at least fair test-retest reliability (i.e., $ICC_{3,1} \geq 0.5$) in all 3 levels of standardized task demands) were considered eligible for the investigation on validity, because it has been defined as a prerequisite to ascertain the test-retest reliability before further exploring the validity of vm-HRV to monitor ITL during exergame-based motor-cognitive training to ensure that observed changes in the variable of interest are attributable to real changes rather than measurement error. Validity was checked by assessing whether there was an effect of game level (i.e. 'easy' vs. 'challenging' vs. 'excessive') on vm-HRV in the 2-way repeated-measures ANOVA (section 'Test-retest reliability of vm-HRV'). In case of a significant main effect of game level and no significant interaction effect, post-hoc tests were computed by calculating pairwise t-test with Bonferroni correction or a Wilcoxon signed-rank test in case of data violating assumptions for parametric analyses [70]. Effect sizes r were calculated for the pairwise comparisons [65, 71] and interpreted as small ($0.1 \leq r < 0.3$), medium ($0.3 \leq r < 0.5$) or large ($r > 0.5$) [70]. To verify that the predetermined levels of external task demands changed ITL (approximated by means of the NASA-TLX rating), the same statistics were computed for the NASA-TLX score.

Sample size justification

Sample size justification was based on the estimation approach for determining sample size for estimating ICC's of Borg et al. [72] derived from Bonnet [73]. The latter provides sample size requirements for estimating ICC's with a desired precision [73] while incorporating Bonett's correction factor. Considering the criterion for good test-retest reliability ($ICC \geq 0.75$) as anticipated ICC and a desired width of the confidence interval of ≤ 0.3 with 50% probability of obtaining the desired precision, a minimum sample size of $n = 38$ was required at a 95% confidence interval.

Results

Recruitment and participant flow

A summary of the participant flow through the study is illustrated in Fig. 2.

Further recruitment was stopped when the planned minimum sample size of 38 participants completed the study.

Baseline data and descriptive statistics

The baseline factors of the study participants are summarized in Table 2.

The subjective ratings of task demands are summarized in Table 3. Significant ($p < 0.001$) differences with large effect sizes (mean $r = 0.980$, range = 0.513–1.229) were observed between all standardized levels of external task demands in both experimental phases, except for the game 'Habitats' level 2 vs. level 3 in experimental phase 1 and the game 'Habitats' level 1 vs. level 2 and level 2 vs. level 3 in experimental phase 2.

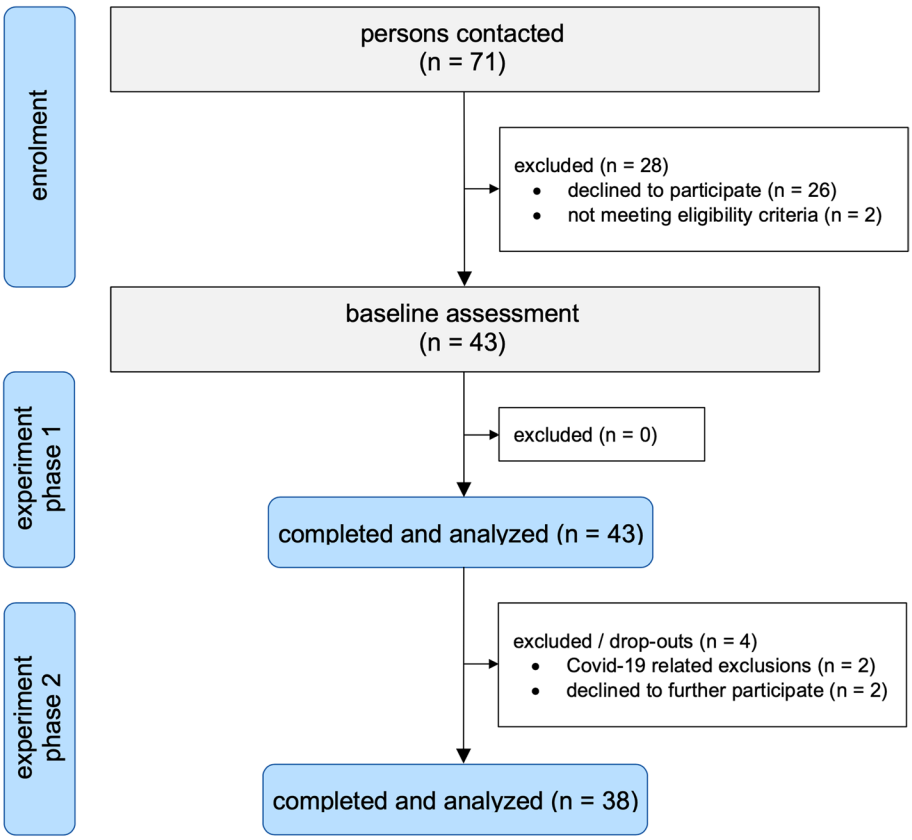

Fig. 2 Summary of the participant flow throughout the study

Table 2 Demographic characteristics of the study population

	Total Sample (*n* = 43)
Age [years]	67.0 (7.0)
Sex [% females]	58.1 (25 females, 18 males)
Body mass index [kg·m^{-2}]	23.7 ± 3.0
Physical Activity [min · week^{-1}]	330 (260)
Intake of cardioactive medication [% of participants]	34.9
Qmci total score	79.1 ± 8.5

Data is reported as mean ± standard deviation for data fulfilling all the assumptions that would subsequently justify parametric statistical analyses and median (interquartile range) for data violating these assumptions

Abbreviations: *Qmci* Quick mild cognitive impairment screen

The mean stepping frequencies to achieve the target of ≥ 40% HRR were 149.5 ± 26.8 and 150.5 ± 25.5 at test and re-test measurement, respectively.

Test-retest reliability of vm-HRV

Table 4 presents the results on test-retest reliability of vm-HRV. The parameters mRR and PNS-Index showed no systematic error and mostly good to excellent relative test-retest reliability (mean $ICC_{3,1}$ = 0.855 (range = 0.434 to 0.939) and 0.787 (range = 0.519 to 0.903)), irrespective of the type of neurocognitive domain trained, experimental phase, or standardized level of external task demands. The mean SEM% and SDD% were 36.4% (range = 24.7 to 75.2) and 100.7% (range = 68.5 to 208.5) for the mRR, and 44.6% (range = 29.0 to 69.4) and 123.7% (range = 80.3 to 192.2) for the PNS-Index, respectively. The remaining vm-HRV parameters mostly showed no systematic error, but inconsistent or poor test-retest reliability.

Validity of vm-HRV to monitor ITL

Table 4 presents the results on the validity of vm-HRV to monitor ITL. The parameters mRR and PNS-Index showed main effects for level with significant ($p < 0.001$) differences and with mostly large effect sizes (mean $r = 0.847$ (range = 0.207 to 1.229)) between all standardized levels of external task demands, irrespective of the type of neurocognitive domain trained and experimental phase. The remaining vm-HRV parameters showed varying results depending on the experimental phase and level of external task demand. In experimental phase 1, there was a main effect for level for the parameters RMSSD and SD1 in the game training attention (game = "Simple"), with significant ($p < 0.001$)

Table 3 Descriptive statistics and analysis of variance (ANOVA) for NASA-TLX total scores

NASA-TLX total score	Experimental Phase 1									Experimental Phase 2								
	Descriptives[a]	Main Effect for Time			Main Effect for Level					Descriptives[a]	Main Effect for Time			Main Effect for Level				
	mean ± SD or median (IQR)	F^b	p^c		F^d	p^e	post-hoc tests			mean ± SD or median (IQR)	F^b	p^c		F^d	p^e	post-hoc tests		
Game: Simple		0.3	0.601	$n=43$	130.5	< 0.001	comparison:	*p*-value:	effect size:		3.1	0.079	$n=35$	51.5	< 0.001	comparison:	*p*-value:	effect size:
Level 1	9.5 (19.0)						Level 1 vs. 2	< 0.001	0.909	37.5 (29.0)						Level 1 vs. 2	< 0.001	0.823
Level 2	16.0 (24.0)						Level 1 vs. 3	< 0.001	1.187	44.5 (35.0)						Level 1 vs. 3	< 0.001	1.164
Level 3	27.5 (26.5)						Level 2 vs. 3	< 0.001	1.133	57.5 (33.5)						Level 2 vs. 3	< 0.001	1.006
Game: Targets		10.3	0.001	$n=42$	132.0	< 0.001	comparison:	*p*-value:	effect size:		6.3	0.012	$n=34$	89.4	< 0.001	comparison:	*p*-value:	effect size:
Level 1	12.0 (21.0)						Level 1 vs. 2	< 0.001	1.168	30.0 (26.5)						Level 1 vs. 2	< 0.001	1.128
Level 2	29.0 (26.3)						Level 1 vs. 3	< 0.001	1.227	48.5 (28.5)						Level 1 vs. 3	< 0.001	1.229
Level 3	45.0 (31.8)						Level 2 vs. 3	< 0.001	1.194	64.0 (30.5)						Level 2 vs. 3	< 0.001	1.174
Game: Tetris		4.9	0.028	$n=42$	76.8	< 0.001	comparison:	*p*-value:	effect size:		16.3	< 0.001	$n=34$	56.0	< 0.001	comparison:	*p*-value:	effect size:
Level 1	19.0 (23.5)						Level 1 vs. 2	< 0.001	0.906	33.0 (29.3)						Level 1 vs. 2	< 0.001	0.994
Level 2	28.5 (40.3)						Level 1 vs. 3	< 0.001	1.189	45.0 (33.5)						Level 1 vs. 3	< 0.001	1.163
Level 3	58.5 (49.5)						Level 2 vs. 3	< 0.001	1.181	60.5 (32.3)						Level 2 vs. 3	< 0.001	0.917
Game: Simon		9.2	0.002	$n=41$	132.0	< 0.001	comparison:	*p*-value:	effect size:		12.2	< 0.001	$n=33$	68.4	< 0.001	comparison:	*p*-value:	effect size:
Level 1	15.0 (25.0)						Level 1 vs. 2	< 0.001	1.151	44.0 (36.0)						Level 1 vs. 2	< 0.001	0.910
Level 2	32.0 (32.5)						Level 1 vs. 3	< 0.001	1.227	52.0 (35.8)						Level 1 vs. 3	< 0.001	1.201
Level 3	60.0 (41.3)						Level 2 vs. 3	< 0.001	1.201	70.0 (32.5)						Level 2 vs. 3	< 0.001	1.162
Game: Habitats		0.0	0.936	$n=42$	4.9	0.013	comparison:	*p*-value:	effect size:		8.0	0.005	$n=35$	12.8	< 0.001	comparison:	*p*-value:	effect size:
Level 1	13.0 (25.0)						Level 1 vs. 2	< 0.001	0.513	31.0 (34.5)						Level 1 vs. 2	0.058	0.320
Level 2	16.5 (28.0)						Level 1 vs. 3	< 0.001	0.515	32.0 (37.0)						Level 1 vs. 3	< 0.001	0.813
Level 3	16.0 (31.3)						Level 2 vs. 3	0.218	0.190	36.5 (31.0)						Level 2 vs. 3	0.003	0.491

Presented are descriptive statistics, *F*-values, and *p*-values of the test and the retest data in Experimental Phase 1 and 2, respectively

Abbreviations: *HF* Absolute power of the high-frequency (0.15–0.4 Hz; HF) band, *HFnu* Relative power of HF (in normal units), *IQR* Interquartile range, *mRR* Mean R-R time interval, *n* Sample size, *NASA-TLX* NASA task load index, *RMSSD* Root mean square of successive RR interval differences, *SD1* Poincaré plot standard deviation perpendicular to the line of identity, *SD* Standard deviation

[a] Descriptive statistics data is reported as mean ± standard deviation for data fulfilling all the assumptions that would subsequently justify parametric statistical analyses and median (interquartile range) for data violating these assumptions

[b] *F*-value for the main effect of timepoint (test- vs. retest-measurement) from the two-way repeated measures ANOVA

[c] *P*-value for the main effect of timepoint (test- vs. retest-measurement) from the two-way repeated measures ANOVA

[d] *F*-value for the main effect of game level (Level 1 = easy, Level 2 = challenging, Level 3 = excessive) from the two-way repeated measures ANOVA

[e] *P*-value for the main effect of game level (Level 1 = easy, Level 2 = challenging, Level 3 = excessive) from the two-way repeated measures ANOVA

Table 4 Test–retest reliability of HRV reactivity

Game: Simple

HRV on-task parameter	Experimental Phase 1															Experimental Phase 2													
	Descriptives[a]	Relative Reliability					Validity							Absolute Reliability		Descriptives[a]	Relative Reliability					Validity							Absolute Reliability	
	mean ± SD or median (IQR)	$ICC_{3,1}$ [95% CI]			F^b	p^c	F^e	p^f	post-hoc tests					SEM %	SDD %	mean ± SD or median (IQR)	$ICC_{3,1}$ [95% CI]			F^b	p^c	F^e	p^f	post-hoc tests					SEM %	SDD %
mRR [ms]					0.00	0.968 [n = 33]	60.9	< 0.001	comparison:	p-value:	effect size:								0.30	0.581 [n = 22]	27.1	< 0.001	comparison:	p-value:	effect size:				
Level 1	726.0 (114.5)	0.923	[0.868, 0.956]	[n = 36, d]					Level 1 vs. 2	< 0.001	1.222			27.7	77.0	500.5 (76.0)	0.833	[0.715, 0.905]	[n = 35, d]					Level 1 vs. 2	< 0.001	0.832	40.9	113.3
Level 2	690.0 (103.5)	0.929	[0.878, 0.960]	[n = 37, d]					Level 1 vs. 3	< 0.001	1.229			26.6	73.9	489.0 (67.3)	0.851	[0.748, 0.914]	[n = 34, d]					Level 1 vs. 3	< 0.001	1.208	39.0	107.0
Level 3	646.0 (122.0)	0.931	[0.881, 0.960]	[n = 38, d]					Level 2 vs. 3	< 0.001	1.227			26.3	72.8	475.0 (52.0)	0.857	[0.756, 0.918]	[n = 35, d]					Level 2 vs. 3	< 0.001	0.952	37.8	104.8
RMSSD [ms]					0.00	0.961 [n = 33]	23.9	< 0.001	comparison:	p-value:	effect size:								0.00	0.884 [n = 22]	0.7	0.508	comparison:	p-value:	effect size:				
Level 1	13.2 (11.6)	0.600	[0.387, 0.753]	[n = 36, d]					Level 1 vs. 2	< 0.001	0.899			63.2	175.3	3.8 (1.8)	0.527	[0.279, 0.710]	[n = 35, d]					Level 1 vs. 2	N/A[h]	N/A[h]	68.8	190.6
Level 2	10.2 (8.8)	0.797	[0.667, 0.880]	[n = 37, d]					Level 1 vs. 3	< 0.001	1.010			45.1	124.9	3.7 (2.0)	0.679	[0.491, 0.807]	[n = 34, d]					Level 1 vs. 3	N/A[h]	N/A[h]	56.7	157.0
Level 3	8.7 (6.5)	0.638	[0.437, 0.778]	[n = 38, d]					Level 2 vs. 3	< 0.001	0.751			60.2	166.8	3.5 (1.4)	0.656	[0.455, 0.794]	[n = 35, d]					Level 2 vs. 3	N/A[h]	N/A[h]	58.7	162.6
HF [ms²]					0.00	0.925 [n = 33]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:								0.30	0.598 [n = 22]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:				
Level 1	70.5 (131.3)	0.382	[0.116, 0.597]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]			78.6	217.9	2.0 (2.3)	0.430	[0.158, 0.641]	[n = 35, d]					N/A[g]	N/A[g]	N/A[g]	75.5	209.3
Level 2	16.0 (42.5)	0.737	[0.576, 0.842]	[n = 37, d]					N/A[g]	N/A[g]	N/A[g]			51.3	142.2	1.0 (1.0)	0.523	[0.283, 0.702]	[n = 34, d]					N/A[g]	N/A[g]	N/A[g]	69.1	191.4
Level 3	10.0 (20.0)	0.812	[0.689, 0.889]	[n = 38, d]					N/A[g]	N/A[g]	N/A[g]			43.4	120.2	1.0 (1.0)	0.066	[-0.225, 0.346]	[n = 35, d]					N/A[g]	N/A[g]	N/A[g]	96.4	267.9
HFnu [nu]					0.10	0.736 [n = 33]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:								0.00	0.866 [n = 22]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:				
Level 1	41.7 (31.8)	0.367	[0.099, 0.586]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]			79.6	220.5	39.9 ± 20.0	0.000	[-0.291, 0.291]	[n = 35, d]					N/A[g]	N/A[g]	N/A[g]	100.0	277.2
Level 2	31.7 (31.8)	0.705	[0.531, 0.822]	[n = 37, d]					N/A[g]	N/A[g]	N/A[g]			54.3	150.6	34.05 (28.7)	0.462	[0.206, 0.659]	[n = 34, d]					N/A[g]	N/A[g]	N/A[g]	73.3	203.3
Level 3	32.3 (38.8)	0.585	[0.367, 0.743]	[n = 38, d]					N/A[g]	N/A[g]	N/A[g]			64.4	178.6	42.0 ± 19.8	0.475	[0.218, 0.671]	[n = 35, d]					N/A[g]	N/A[g]	N/A[g]	72.5	200.8
SD1 [ms]					0.00	0.867 [n = 33]	25.4	< 0.001	comparison:	p-value:	effect size:								0.00	0.908 [n = 22]	0.8	0.459	comparison:	p-value:	effect size:				
Level 1	9.4 (8.3)	0.601	[0.387, 0.753]	[n = 36, d]					Level 1 vs. 2	< 0.001	0.895			63.2	175.1	2.8 (1.3)	0.530	[0.282, 0.711]	[n = 35, d]					Level 1 vs. 2	N/A[h]	N/A[h]	68.6	190.0
Level 2	7.3 (6.3)	0.796	[0.665, 0.880]	[n = 37, d]					Level 1 vs. 3	< 0.001	1.048			45.2	125.2	2.6 (1.4)	0.677	[0.487, 0.805]	[n = 34, d]					Level 1 vs. 3	N/A[h]	N/A[h]	56.8	157.5
Level 3	6.1 (4.6)	0.635	[0.433, 0.776]	[n = 38, d]					Level 2 vs. 3	< 0.001	0.784			60.4	167.5	2.5 (1.0)	0.650	[0.447, 0.790]	[n = 35, d]					Level 2 vs. 3	N/A[h]	N/A[h]	59.2	164.0
PNSIndex []					0.00	0.869 [n = 33]	43.8	< 0.001	comparison:	p-value:	effect size:								0.10	0.703 [n = 22]	26.2	< 0.001	comparison:	p-value:	effect size:				
Level 1	-1.7 (0.8)	0.870	[0.781, 0.925]	[n = 36, d]					Level 1 vs. 2	< 0.001	1.092			36.1	99.9	-3.2 (0.6)	0.833	[0.715, 0.905]	[n = 35, d]					Level 1 vs. 2	< 0.001	0.738	40.9	113.3
Level 2	-1.9 (0.9)	0.898	[0.826, 0.941]	[n = 37, d]					Level 1 vs. 3	< 0.001	1.205			31.9	88.5	-3.3 (0.6)	0.808	[0.681, 0.888]	[n = 34, d]					Level 1 vs. 3	< 0.001	1.127	43.8	121.5
Level 3	-2.1 (1.0)	0.898	[0.826, 0.941]	[n = 38, d]					Level 2 vs. 3	< 0.001	1.120			31.9	88.5	-3.4 (0.5)	0.837	[0.724, 0.906]	[n = 35, d]					Level 2 vs. 3	< 0.001	0.928	40.4	111.9

Game: Targets

HRV on-task parameter	Experimental Phase 1															Experimental Phase 2													
	Descriptives[a]	Relative Reliability					Validity							Absolute Reliability		Descriptives[a]	Relative Reliability					Validity							Absolute Reliability	
	mean ± SD or median (IQR)	$ICC_{3,1}$ [95% CI]			F^b	p^c	F^e	p^f	post-hoc tests					SEM %	SDD %	mean ± SD or median (IQR)	$ICC_{3,1}$ [95% CI]			F^b	p^c	F^e	p^f	post-hoc tests					SEM %	SDD %
mRR [ms]					0.80	0.361 [n = 33]	83.0	< 0.001	comparison:	p-value:	effect size:								1.40	0.229 [n = 25]	28.5	< 0.001	comparison:	p-value:	effect size:				
Level 1	727.0 (131.3)	0.929	[0.877, 0.959]	[n = 36, d]					Level 1 vs. 2	< 0.001	1.229			26.6	73.9	507.0 (92.0)	0.833	[0.720, 0.903]	[n = 36, d]					Level 1 vs. 2	< 0.001	1.148	40.9	113.3
Level 2	657.0 (112.0)	0.918	[0.863, 0.952]	[n = 36, d]					Level 1 vs. 3	< 0.001	1.229			28.6	79.4	480.5 (78.5)	0.843	[0.738, 0.908]	[n = 37, d]					Level 1 vs. 3	< 0.001	1.129	39.6	109.8
Level 3	638.1 (119.8)	0.939	[0.894, 0.965]	[n = 36, d]					Level 2 vs. 3	< 0.001	1.203			24.7	68.5	464.0 (77.8)	0.872	[0.784, 0.926]	[n = 38, d]					Level 2 vs. 3	< 0.001	0.921	35.8	99.2
RMSSD [ms]					0.00	0.937 [n = 33]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:								0.60	0.425 [n = 25]	7.6	< 0.001	comparison:	p-value:	effect size:				
Level 1	12.0 (10.7)	0.714	[0.544, 0.828]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]			53.5	148.2	4.0 (2.9)	0.702	[0.523, 0.822]	[n = 36, d]					Level 1 vs. 2	0.018	0.474	54.6	151.3
Level 2	9.7 (7.0)	0.005	[-0.263, 0.272]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			99.7	276.5	3.6 (2.0)	0.827	[0.713, 0.899]	[n = 37, d]					Level 1 vs. 3	< 0.001	0.720	41.6	115.3
Level 3	6.9 (5.7)	0.545	[0.314, 0.715]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]			67.5	187.0	3.6 (2.0)	0.858	[0.761, 0.917]	[n = 38, d]					Level 2 vs. 3	0.047	0.397	37.7	104.5
HF [ms²]					0.10	0.773 [n = 33]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:								0.50	0.459 [n = 25]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:				
Level 1	59.5 (134.0)	0.604	[0.391, 0.755]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]			62.9	174.4	2.0 (6.0)	0.040	[-0.246, 0.319]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]	98.0	271.6
Level 2	16.0 (35.0)	0.174	[-0.098, 0.422]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			90.9	251.9	1.0 (1.0)	0.017	[-0.263, 0.294]	[n = 37, d]					N/A[g]	N/A[g]	N/A[g]	99.1	274.9
Level 3	9.0 (16.0)	0.726	[0.561, 0.835]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]			52.3	145.1	1.0 (1.3)	0.061	[-0.222, 0.334]	[n = 38, d]					N/A[g]	N/A[g]	N/A[g]	96.9	268.6
HFnu [nu]					2.50	0.116 [n = 33]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:								0.00	0.892 [n = 25]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:				
Level 1	43.7 ± 20.4	0.418	[0.157, 0.624]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]			76.3	211.5	39.6 (33.6)	0.153	[-0.135, 0.418]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]	92.0	255.1
Level 2	26.3 (25.6)	0.562	[0.347, 0.721]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			66.2	183.4	27.0 (34.0)	0.706	[0.533, 0.823]	[n = 37, d]					N/A[g]	N/A[g]	N/A[g]	54.2	150.3
Level 3	28.2 (24.4)	0.614	[0.405, 0.762]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]			62.4	172.2	39.0 ± 19.9	0.397	[0.133, 0.608]	[n = 38, d]					N/A[g]	N/A[g]	N/A[g]	77.7	215.2
SD1 [ms]					0.00	0.945 [n = 33]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:								0.60	0.422 [n = 25]	7.7	< 0.001	comparison:	p-value:	effect size:				
Level 1	8.6 (7.7)	0.715	[0.545, 0.829]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]			53.4	148.0	2.8 (2.1)	0.697	[0.516, 0.818]	[n = 36, d]					Level 1 vs. 2	0.009	0.522	55.0	152.6
Level 2	6.9 (4.9)	0.005	[-0.262, 0.272]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			99.7	276.5	2.6 (1.4)	0.827	[0.713, 0.899]	[n = 37, d]					Level 1 vs. 3	< 0.001	0.711	41.6	115.3
Level 3	4.9 (4.1)	0.545	[0.314, 0.715]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]			67.5	187.0	2.5 (1.4)	0.861	[0.766, 0.919]	[n = 38, d]					Level 2 vs. 3	0.057	0.381	37.3	103.3
PNSIndex []					0.70	0.391 [n = 33]	80.8	< 0.001	comparison:	p-value:	effect size:								1.90	0.172 [n = 25]	29.9	< 0.001	comparison:	p-value:	effect size:				
Level 1	-1.7 (0.9)	0.903	[0.834, 0.944]	[n = 36, d]					Level 1 vs. 2	< 0.001	1.167			31.1	86.3	-3.1 (0.8)	0.816	[0.693, 0.893]	[n = 36, d]					Level 1 vs. 2	< 0.001	1.161	42.9	119.0
Level 2	-2.1 (0.9)	0.619	[0.422, 0.761]	[n = 39, d]					Level 1 vs. 3	< 0.001	1.227			61.7	171.1	-3.4 (0.7)	0.856	[0.758, 0.916]	[n = 37, d]					Level 1 vs. 3	< 0.001	1.113	37.9	105.2
Level 3	-2.4 (0.8)	0.885	[0.805, 0.934]	[n = 36, d]					Level 2 vs. 3	< 0.001	1.182			33.9	94.0	-3.5 (0.7)	0.893	[0.817, 0.938]	[n = 38, d]					Level 2 vs. 3	< 0.001	0.893	32.7	90.7

Game: Tetris

HRV on-task parameter	Experimental Phase 1															Experimental Phase 2													
	Descriptives[a]	Relative Reliability					Validity							Absolute Reliability		Descriptives[a]	Relative Reliability					Validity							Absolute Reliability	
	mean ± SD or median (IQR)	$ICC_{3,1}$ [95% CI]			F^b	p^c	F^e	p^f	post-hoc tests					SEM %	SDD %	mean ± SD or median (IQR)	$ICC_{3,1}$ [95% CI]			F^b	p^c	F^e	p^f	post-hoc tests					SEM %	SDD %
mRR [ms]					0.00	0.893 [n = 33]	41.4	< 0.001	comparison:	p-value:	effect size:								2.90	0.087 [n = 22]	21.1	< 0.001	comparison:	p-value:	effect size:				
Level 1	723.0 (124.0)	0.886	[0.811, 0.933]	[n = 39, d]					Level 1 vs. 2	< 0.001	0.624			33.8	93.6	530.0 (76.0)	0.883	[0.800, 0.933]	[n = 36, d]					Level 1 vs. 2	< 0.001	0.849	34.2	94.8
Level 2	711.0 (119.0)	0.910	[0.849, 0.947]	[n = 39, d]					Level 1 vs. 3	< 0.001	1.161			30.0	83.2	518.0 (60.0)	0.847	[0.740, 0.912]	[n = 33, d]					Level 1 vs. 3	< 0.001	1.227	39.1	108.4
Level 3	687.0 (116.0)	0.899	[0.831, 0.941]	[n = 38, d]					Level 2 vs. 3	< 0.001	0.984			31.8	88.1	503.0 (74.0)	0.848	[0.742, 0.913]	[n = 34, d]					Level 2 vs. 3	< 0.001	0.933	39.0	108.1
RMSSD [ms]					0.00	0.865 [n = 33]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:								0.10	0.700 [n = 22]	4.1	0.018	comparison:	p-value:	effect size:				
Level 1	9.6 (9.2)	0.086	[-0.186, 0.345]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			95.6	265.0	4.1 (3.0)	0.707	[0.530, 0.825]	[n = 36, d]					Level 1 vs. 2	0.708	0.080	54.1	150.0
Level 2	10.0 (8.8)	0.807	[0.689, 0.884]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			43.9	121.8	4.1 (3.6)	0.621	[0.407, 0.771]	[n = 33, d]					Level 1 vs. 3	0.096	0.355	61.6	170.6
Level 3	9.2 (8.5)	0.677	[0.498, 0.801]	[n = 38, d]					N/A[g]	N/A[g]	N/A[g]			56.8	157.5	3.7 (2.4)	0.715	[0.539, 0.831]	[n = 34, d]					Level 2 vs. 3	0.196	0.276	53.4	148.0
HF [ms²]					0.40	0.552 [n = 33]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:								0.00	0.930 [n = 22]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:				
Level 1	33.0 (51.0)	0.028	[-0.241, 0.293]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			98.6	273.3	3.0 (5.0)	0.341	[0.064, 0.569]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]	81.2	225.0
Level 2	33.0 (73.0)	0.571	[0.358, 0.727]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			65.5	181.6	2.0 (5.0)	0.077	[-0.215, 0.356]	[n = 33, d]					N/A[g]	N/A[g]	N/A[g]	96.1	266.3
Level 3	28.0 (44.0)	0.397	[0.142, 0.603]	[n = 38, d]					N/A[g]	N/A[g]	N/A[g]			77.7	215.2	2.0 (4.5)	0.407	[0.136, 0.621]	[n = 34, d]					N/A[g]	N/A[g]	N/A[g]	77.0	213.5
HFnu [nu]					0.00	0.878 [n = 33]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:								1.60	0.210 [n = 22]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:				
Level 1	20.8 (18.5)	0.237	[-0.033, 0.474]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			87.3	242.1	28.4 (28.8)	0.259	[-0.025, 0.505]	[n = 36, d]					N/A[g]	N/A[g]	N/A[g]	86.1	238.6
Level 2	19.5 (20.0)	0.385	[0.131, 0.591]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			78.4	217.4	23.2 (24.0)	0.355	[0.075, 0.582]	[n = 33, d]					N/A[g]	N/A[g]	N/A[g]	80.3	222.6
Level 3	21.2 (21.0)	0.590	[0.380, 0.743]	[n = 38, d]					N/A[g]	N/A[g]	N/A[g]			64.0	177.5	27.1 (24.4)	0.681	[0.490, 0.810]	[n = 34, d]					N/A[g]	N/A[g]	N/A[g]	56.5	156.6
SD1 [ms]					0.00	0.882 [n = 33]	N/A[i]	N/A[i]	comparison:	p-value:	effect size:								0.10	0.745 [n = 22]	4.0	0.019	comparison:	p-value:	effect size:				
Level 1	6.8 (6.6)	0.087	[-0.185, 0.346]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			95.6	264.9	2.9 (2.2)	0.707	[0.530, 0.825]	[n = 36, d]					Level 1 vs. 2	0.695	0.084	54.1	150.0
Level 2	7.1 (6.3)	0.808	[0.689, 0.884]	[n = 39, d]					N/A[g]	N/A[g]	N/A[g]			43.8	121.5	2.9 (2.6)	0.623	[0.409, 0.772]	[n = 33, d]					Level 1 vs. 3	0.107	0.344	61.4	170.2
Level 3	6.6 (6.0)	0.677	[0.497, 0.801]	[n = 38, d]					N/A[g]	N/A[g]	N/A[g]			56.8	157.5	2.6 (1.8)	0.710	[0.531, 0.828]	[n = 34, d]					Level 2 vs. 3	0.154	0.304	53.9	149.3
PNSIndex []					0.00	0.924 [n = 33]	17.2	< 0.001	comparison:	p-value:	effect size:								2.00	0.162 [n = 22]	17.4	< 0.001	comparison:	p-value:	effect size:				
Level 1	-1.87 (0.7)	0.644	[0.455, 0.778]	[n = 39, d]					Level 1 vs. 2	0.030	0.357			59.7	165.4	-2.9 (0.6)	0.861	[0.765, 0.920]	[n = 36, d]					Level 1 vs. 2	< 0.001	0.728	37.3	103.3
Level 2	-1.91 (0.8)	0.898	[0.829, 0.939]	[n = 39, d]					Level 1 vs. 3	< 0.001	0.899			31.9	88.5	-3.0 (0.6)	0.790	[0.651, 0.878]	[n = 33, d]					Level 1 vs. 3	< 0.001	1.134	45.8	127.0
Level 3	-2.0 (0.8)	0.845	[0.745, 0.908]	[n = 38, d]					Level 2 vs. 3	< 0.001	0.657			39.4	109.1	-3.2 (0.7)	0.852	[0.747, 0.915]	[n = 34, d]					Level 2 vs. 3	< 0.001	0.835	38.5	106.6

differences and large effect sizes (ranging between 0.751 and 1.048) between all standardized levels of external task demands. In experimental phase 2 there was a main effect for level for the game training attention (game = "Simple") as well as learning and memory (game = "Simon"), whereas there was no significant main effect for level for the game training visuospatial skills (game = "Tetris"), and inconsistent findings for the games training executive functions (games = "Targets" and "Habitats").

Table 4 (continued)

Game: Simon

HRV on-task parameter	Experimental Phase 1													Experimental Phase 2											
	Descriptives[a]	Relative Reliability			Validity								Absolute Reliability		Descriptives[a]	Relative Reliability			Validity						Absolute Reliability	
	mean ± SD or median (IQR)	$ICC_{3,1}$ [95% CI]			F^b	p^c		F^e	p^f	post-hoc tests			SEM %	SDD %	mean ± SD or median (IQR)	$ICC_{3,1}$ [95% CI]			F^e	p^f	post-hoc tests			SEM %	SDD %
mRR [ms]					0.1	0.716	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size							1.4	0.229	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size		
Level 1	692.0 (92.5)	0.907	[0.843, 0.946]	ns, d						N/A[g]	N/A[g]	N/A[g]	30.5	84.5	504.0 (75.5)	0.840	[0.724, 0.910]	ns, d						N/A[g]	N/A[g]	N/A[g]	40.0	110.9
Level 2	677.0 (104.5)	0.457	[0.204, 0.652]	ns, d						N/A[g]	N/A[g]	N/A[g]	73.7	204.3	502.0 (76.0)	0.842	[0.734, 0.908]	ns, d						N/A[g]	N/A[g]	N/A[g]	39.7	110.2
Level 3	697.0 (120.5)	0.916	[0.859, 0.951]	ns, d						N/A[g]	N/A[g]	N/A[g]	30.0	80.3	515.0 (76.0)	0.434	[0.168, 0.641]	ns, d						N/A[g]	N/A[g]	N/A[g]	75.2	208.5
RMSSD [ms]					1.6	0.209	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size							1.6	0.207	ns, d	0.1	0.898	comparison	p-value	effect size		
Level 1	10.3 (8.1)	0.036	[-0.241, 0.308]	ns, d						N/A[g]	N/A[g]	N/A[g]	98.2	272.2	3.7 (2.0)	0.762	[0.602, 0.863]	ns, d						Level 1 vs. 2	N/A[h]	N/A[h]	48.8	135.2
Level 2	10.5 (8.1)	0.668	[0.478, 0.798]	ns, d						N/A[g]	N/A[g]	N/A[g]	57.6	159.7	3.7 (2.1)	0.740	[0.578, 0.845]	ns, d						Level 1 vs. 3	N/A[h]	N/A[h]	51.0	141.3
Level 3	11.8 (9.6)	0.027	[-0.245, 0.296]	ns, d						N/A[g]	N/A[g]	N/A[g]	98.6	273.4	3.7 (2.0)	0.896	[0.819, 0.941]	ns, d						Level 2 vs. 3	N/A[h]	N/A[h]	32.2	89.4
HF [ms²]					1.9	0.173	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size							1.4	0.242	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size		
Level 1	30.5 (52.8)	0.028	[-0.248, 0.300]	ns, d						N/A[g]	N/A[g]	N/A[g]	98.6	273.3	2.0 (2.0)	0.701	[0.512, 0.826]	ns, d						N/A[g]	N/A[g]	N/A[g]	54.7	151.6
Level 2	50.0 (91.0)	0.674	[0.487, 0.802]	ns, d						N/A[g]	N/A[g]	N/A[g]	57.1	158.3	1.0 (3.0)	0.281	[-0.002, 0.522]	ns, d						N/A[g]	N/A[g]	N/A[g]	84.8	235.0
Level 3	30.0 (61.5)	0.145	[-0.131, 0.401]	ns, d						N/A[g]	N/A[g]	N/A[g]	92.5	256.3	2.0 (2.0)	0.851	[0.746, 0.915]	ns, d						N/A[g]	N/A[g]	N/A[g]	38.6	107.0
HFnu [nu]					0.0	0.945	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size							2.7	0.101	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size		
Level 1	40.0 (35.6)	0.476	[0.231, 0.664]	ns, d						N/A[g]	N/A[g]	N/A[g]	72.4	200.6	39.1 ± 21.4	0.503	[0.243, 0.695]	ns, d						N/A[g]	N/A[g]	N/A[g]	70.5	195.4
Level 2	43.9 ± 22.8	0.541	[0.309, 0.712]	ns, d						N/A[g]	N/A[g]	N/A[g]	67.7	187.8	33.7 (33.1)	0.177	[-0.111, 0.438]	ns, d						N/A[g]	N/A[g]	N/A[g]	90.7	251.5
Level 3	14.9 (16.1)	0.557	[0.336, 0.719]	ns, d						N/A[g]	N/A[g]	N/A[g]	66.6	184.5	24.6 (31.3)	0.140	[-0.153, 0.410]	ns, d						N/A[g]	N/A[g]	N/A[g]	92.7	257.1
SD1 [ms]					1.6	0.212	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size							1.7	0.199	ns, d	0.1	0.848	comparison	p-value	effect size		
Level 1	7.3 (5.7)	0.037	[-0.240, 0.309]	ns, d						N/A[g]	N/A[g]	N/A[g]	98.1	272.0	2.6 (1.4)	0.761	[0.600, 0.862]	ns, d						Level 1 vs. 2	N/A[h]	N/A[h]	48.9	135.5
Level 2	7.5 (5.8)	0.669	[0.480, 0.799]	ns, d						N/A[g]	N/A[g]	N/A[g]	57.5	159.5	2.6 (1.5)	0.741	[0.579, 0.846]	ns, d						Level 1 vs. 3	N/A[h]	N/A[h]	50.9	141.1
Level 3	8.4 (6.8)	0.027	[-0.246, 0.296]	ns, d						N/A[g]	N/A[g]	N/A[g]	98.6	273.4	2.6 (1.5)	0.897	[0.822, 0.942]	ns, d						Level 2 vs. 3	N/A[h]	N/A[h]	32.1	89.0
PNSIndex []					0.2	0.646	ns, d	6.0	0.004	comparison	p-value	effect size							2.4	0.123	ns, d	3.2	0.042	comparison	p-value	effect size		
Level 1	-1.9 (0.9)	0.562	[0.340, 0.725]	ns, d						Level 1 vs. 2	< 0.001	1.095	66.2	183.4	-3.1 (0.7)	0.805	[0.669, 0.889]	ns, d						Level 1 vs. 2	0.280	0.220	44.2	122.4
Level 2	-1.9 (0.9)	0.916	[0.857, 0.952]	ns, d						Level 1 vs. 3	0.002	0.533	29.0	80.3	-3.2 (0.6)	0.805	[0.676, 0.886]	ns, d						Level 1 vs. 3	0.185	0.270	44.2	122.4
Level 3	-1.9 (0.8)	0.519	[0.289, 0.693]	ns, d						Level 2 vs. 3	0.087	0.293	69.4	192.2	-3.1 (0.6)	0.822	[0.700, 0.897]	ns, d						Level 2 vs. 3	0.005	0.576	42.2	116.9

Game: Habitats

HRV on-task parameter	Experimental Phase 1													Experimental Phase 2											
	Descriptives[a]	Relative Reliability			Validity								Absolute Reliability		Descriptives[a]	Relative Reliability			Validity						Absolute Reliability	
	mean ± SD or median (IQR)	$ICC_{3,1}$ [95% CI]			F^b	p^c		F^e	p^f	post-hoc tests			SEM %	SDD %	mean ± SD or median (IQR)	$ICC_{3,1}$ [95% CI]			F^e	p^f	post-hoc tests			SEM %	SDD %
mRR [ms]					3.4	0.067	ns, d	4.1	0.022	comparison	p-value	effect size							0.2	0.626	ns, d	7.2	0.001	comparison	p-value	effect size		
Level 1	712.0 (118.3)	0.925	[0.872, 0.957]	ns, d						Level 1 vs. 2	0.257	0.207	27.4	75.9	513.0 (83.5)	0.890	[0.814, 0.936]	ns, d						Level 1 vs. 2	0.014	0.475	33.2	91.9
Level 2	706.0 (109.0)	0.873	[0.783, 0.927]	ns, d						Level 1 vs. 3	< 0.001	0.655	35.6	98.8	511.0 (78.3)	0.891	[0.814, 0.937]	ns, d						Level 1 vs. 3	< 0.001	0.705	33.0	91.5
Level 3	708.5 (113.0)	0.893	[0.819, 0.938]	ns, d						Level 2 vs. 3	0.005	0.518	32.7	90.7	499.0 (74.0)	0.861	[0.765, 0.920]	ns, d						Level 2 vs. 3	0.015	0.469	37.3	103.3
RMSSD [ms]					3.9	0.049	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size							0.1	0.802	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size		
Level 1	11.8 (8.8)	0.632	[0.432, 0.772]	ns, d						N/A[g]	N/A[g]	N/A[g]	60.7	168.1	3.8 (3.5)	0.266	[-0.009, 0.504]	ns, d						N/A[g]	N/A[g]	N/A[g]	85.7	237.5
Level 2	11.7 (8.4)	0.044	[-0.242, 0.322]	ns, d						N/A[g]	N/A[g]	N/A[g]	97.8	271.0	3.9 (2.6)	0.245	[-0.036, 0.490]	ns, d						N/A[g]	N/A[g]	N/A[g]	86.9	240.8
Level 3	12.2 (10.5)	0.059	[-0.220, 0.328]	ns, d						N/A[g]	N/A[g]	N/A[g]	97.0	268.9	3.6 (2.4)	0.654	[0.455, 0.790]	ns, d						N/A[g]	N/A[g]	N/A[g]	58.8	163.0
HF [ms²]					4.7	0.030	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size							0.4	0.527	ns, d	2.7	0.074	comparison	p-value	effect size		
Level 1	45.5 (74.8)	0.505	[0.268, 0.685]	ns, d						N/A[g]	N/A[g]	N/A[g]	70.4	195.0	2.0 (3.5)	0.858	[0.764, 0.917]	ns, d						Level 1 vs. 2	N/A[h]	N/A[h]	37.7	104.5
Level 2	47.0 (74.0)	0.000	[-0.283, 0.283]	ns, d						N/A[g]	N/A[g]	N/A[g]	100.0	277.2	2.0 (2.0)	0.723	[0.557, 0.834]	ns, d						Level 1 vs. 3	N/A[h]	N/A[h]	52.6	145.9
Level 3	51.5 (97.5)	0.003	[-0.272, 0.278]	ns, d						N/A[g]	N/A[g]	N/A[g]	99.8	276.8	2.0 (2.0)	0.694	[0.512, 0.817]	ns, d						Level 2 vs. 3	N/A[h]	N/A[h]	55.3	153.3
HFnu [nu]					1.9	0.164	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size							0.1	0.743	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size		
Level 1	41.2 ± 20.9	0.336	[0.068, 0.559]	ns, d						N/A[g]	N/A[g]	N/A[g]	81.5	225.9	37.1 ± 16.8	0.458	[0.209, 0.651]	ns, d						N/A[g]	N/A[g]	N/A[g]	73.6	204.1
Level 2	29.3 (31.9)	0.467	[0.213, 0.662]	ns, d						N/A[g]	N/A[g]	N/A[g]	73.0	202.4	32.8 (30.2)	0.207	[-0.076, 0.459]	ns, d						N/A[g]	N/A[g]	N/A[g]	89.1	246.8
Level 3	32.0 (39.3)	0.547	[0.320, 0.714]	ns, d						N/A[g]	N/A[g]	N/A[g]	67.3	186.6	39.1 ± 19.0	0.490	[0.241, 0.679]	ns, d						N/A[g]	N/A[g]	N/A[g]	71.4	198.0
SD1 [ms]					3.9	0.048	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size							0.1	0.792	ns, d	N/A[g]	N/A[g]	comparison	p-value	effect size		
Level 1	8.4 (6.4)	0.632	[0.432, 0.772]	ns, d						N/A[g]	N/A[g]	N/A[g]	60.7	168.1	2.7 (2.5)	0.267	[-0.009, 0.504]	ns, d						N/A[g]	N/A[g]	N/A[g]	85.6	237.3
Level 2	8.3 (5.9)	0.044	[-0.242, 0.322]	ns, d						N/A[g]	N/A[g]	N/A[g]	97.8	271.0	2.8 (1.8)	0.244	[-0.037, 0.490]	ns, d						N/A[g]	N/A[g]	N/A[g]	86.9	241.0
Level 3	8.7 (7.5)	0.059	[-0.220, 0.328]	ns, d						N/A[g]	N/A[g]	N/A[g]	97.0	268.9	2.6 (1.7)	0.656	[0.459, 0.792]	ns, d						N/A[g]	N/A[g]	N/A[g]	58.7	162.6
PNSIndex []					4.3	0.037	ns, d	3.8	0.024	comparison	p-value	effect size							0.1	0.787	ns, d	4.7	0.010	comparison	p-value	effect size		
Level 1	-1.7 (0.9)	0.843	[0.740, 0.908]	ns, d						Level 1 vs. 2	0.011	0.464	39.6	109.8	-3.0 (0.7)	0.705	[0.533, 0.820]	ns, d						Level 1 vs. 2	0.027	0.425	54.3	150.6
Level 2	-1.8 (0.8)	0.548	[0.315, 0.719]	ns, d						Level 1 vs. 3	< 0.001	0.642	67.2	186.4	-3.4 (0.6)	0.654	[0.458, 0.789]	ns, d						Level 1 vs. 3	< 0.001	0.658	58.8	163.0
Level 3	-1.7 (0.7)	0.585	[0.370, 0.741]	ns, d						Level 2 vs. 3	0.151	0.262	64.4	178.6	-3.1 (0.7)	0.824	[0.705, 0.897]	ns, d						Level 2 vs. 3	0.037	0.401	42.0	116.3

Presented are descripitve statistics, ICCs, *F*-values, *p*-values, SEM%, and SDD%, of the test and the retest data in Experimental Phase 1 and 2, respectively

Abbreviations: *HF* Absolute power of the high-frequency (0.15–0.4 Hz; HF) band, *HFnu* Relative power of HF (in normal units), *ICC* Intraclass Correlation Coefficients, *IQR* Interquartile range, *mRR* Mean R-R time interval, *n* Sample size, *NASA-TLX* NASA task load index, *RMSSD* Root mean square of successive RR interval differences, *SD1* Poincaré plot standard deviation perpendicular to the line of identity, *SD* Standard deviation, *SDD* Smallest detectable difference, *SEM* Standard error of measurement

[a] Descriptive statistics; data is reported as mean ± standard deviation for data fulfilling all the assumptions that would subsequently justify parametric statistical analyses and median (interquartile range) for data violating these assumptions

[b] *F*-value for the main effect of timepoint (test- vs. retest-measurement) from the two-way repeated measures ANOVA

[c] *P*-value for the main effect of timepoint (test- vs. retest-measurement) from the two-way repeated measures ANOVA

[d] Missing data due to low quality data (≥ 5% of beats corrected by the automatic beat correction algorithm of Kubios HRV Premium) that was excluded from analysis

[e] *F*-value for the main effect of game level (Level 1 = easy, Level 2 = challenging, Level 3 = excessive) from the two-way repeated measures ANOVA

[f] *P*-value for the main effect of game level (Level 1 = easy, Level 2 = challenging, Level 3 = excessive) from the two-way repeated measures ANOVA

[g] Not applicable, because criteria for analysis were not met. In particular: Only outcome measures with at least a fair test-retest reliability (i.e., ICC3,1 ≥ 0.5) in all three levels of standardized task demands (i.e., Level 1 = easy, Level 2 = challenging, and Level 3 = excessive) were considered eligible for the exploration on validity

[h] Not applicable, because main effect is not significant = > no pairwise post-hoc comparisons

Colour coding for the interpretation of test-retest reliability:

ICC ≥ 0.9: excellent | 0.75 ≤ ICC < 0.9: good | 0.5 ≤ ICC < 0.75: fair | ICC < 0.50: poor

Discussion

This study evaluated the test-retest reliability of vm-HRV during exergame-based motor-cognitive training in relation to different exergame demands in HOA and its validity as a biomarker to monitor ITL. The results revealed the following key findings: The mRR and PNS-Index, measured with the Polar H10 sensor and calculated with Kubios HRV Premium, showed mostly good to excellent test-retest reliability without systematic error and consistent differences between the standardized levels of external task demands with mostly large effect sizes. These findings persisted irrespective of the type of

neurocognitive domain trained and when only the game demands (motoric and cognitive demands) were manipulated while the physical intensity was kept constant at a moderate level. The remaining vm-HRV parameters showed inconsistent results or poor test-retest reliability and validity.

Test-retest reliability of vm-HRV

To the best of our knowledge, there is only one comparable study that investigated test-retest reliability of HRV during physical or cognitive exercise in relation to different task demands in HOA. Mukherjee et al. [74] measured HRV with on-lead portable electrocardiogram (ECG) during 2 visual working memory tasks requiring different levels of mental effort (i.e., an "easy" and a "hard" trial) in 40 healthy older adults (age = 73.1 ± 4.9 years). They found high test-retest reliability for most HRV parameters, while time domain measures were the most reliable in both task conditions with Kendall's τ ranging from 0.26 to 0.74 [74]. Another study by Guijt et al. [75] found good test-retest reliability during laboratory cycling at light exercise intensity in 26 healthy adults (age = 29.8 ± 8.5 years) in the parameters SDNN (ICC = 0.85 (0.70 – 0.93)) and RMSSD (ICC = 0.84 (0.67 – 0.92)), measured with a one-lead ECG via portable heart rate monitor.

Our results for raw data directly exported from Polar (mRR) and the PNS-Index are consistent with these findings. In contrast to the findings of these 2 studies, however, we had mixed findings for the reliability of remaining vm-HRV parameters. These inconsistent findings might be a result of the differing population characteristics (i.e., healthy adults analyzed in Guijt et al. [75] versus HOA analyzed in Mukherjee et al. [74] and this study) as well as measurement methodologies (i.e., different recording devices and durations that included Co2ntrol (Decon Medical Systems, Weesp, The Netherlands) [75], one-lead ECG with BioSemi (Bio- Semi, Amsterdam, The Netherlands) [74], or Polar HR monitor (Polar M430) and sensor (Polar H10) (this study) and durations of measurement of 10 min [75], 5 min [74], or 1 min (this study)) and conditions (i.e., different levels of physical and cognitive task demands as well as targeted neurocognitive domains (as defined in [76] in line with the Diagnostic and Statistical Manual of Mental Disorders 5th Edition (DSM-5) [77]) [37].

Regarding the measurement methodologies and conditions, Board et al. [36] systematically reviewed the literature and found excellent agreement of mRR measured with different Polar devices in different body positions at rest (ICCs between 0.94 and 1.00) as well as during exercise (ICCs between 0.93 (at vigorous exercise intensity) to 1.00 (at light exercise intensity) compared to multi-lead ECG. Additionally, they concluded that raw data on inter-beat interval time series derived from Polar heart rate monitors are valid for subsequent HRV analysis using validated Kubios HRV software [36], and that the calculated HRV parameters were found to be interchangeable when comparing values derived from the HR monitor-derived time series data with those derived from the multi-lead ECG data. [36]. Because we used the validated automatic beat correction algorithm and noise handling provided by Kubios HRV Premium [55], our mixed findings are likely to be primarily related to the short measurement duration and the high inter-individual variability of vm-HRV [78].

In contrast to Mukherjee et al. (2011) and Guijt et al. (2007), who used measurement durations of 5 min [74] and 10 min [75], we had a shorter measurement duration of only 1 min. Differences in contextual factors (such as age, health, recording methods, measurement conditions, artifacting procedures) may have greater impact on ultra-short-term measurements (< 5 min of measurement) than on longer recordings [37]. To check whether the inter-individual variability can be reduced, we repeated all analyses for vm-HRV reactivity (i.e., the absolute change from resting-state to on-task) [23], but did not find any relevant improvements compared to our original analyses (Supplementary File 4). Therefore, although it has been reported that measurements as short as 1 min may be sufficient to measure resting HR, SDNN and RMSSD [37], our data indicate that a measurement duration of 1 min may be too short for reliable measurement of RMSSD, HF, HFnu, and SD1 during exergame-based training or motor-cognitive training in general.

Validity of vm-HRV to monitor ITL

To the best of our knowledge, this is the first study investigating phasic vm-HRV responses to exergaming. Nevertheless, our findings are consistent with recent literature regarding the sensitivity of vm-HRV to changes in task load of simultaneous motor-cognitive exercises and serious gaming in HOA. Hou et al. [79] analyzed HRV responses to serious games in HOA. They found significant decreases in vm-HRV during serious gaming, which differed between a cognitive aptitude game and reaction time games [79]. They replicated these findings in a consecutive study that also included patients with mild cognitive impairment. Significant decreases in vm-HRV were found during serious gaming, whereas the cognitive status of the study participants had no significant effect on the HRV [80]. This is consistent with the results of a systematic review that found significant withdrawal of vm-HRV in HOA during cognitive exercises, but contradicts the finding that vm-HRV in cognitive tasks is

dependent on participant characteristics (i.e. level of cognitive functioning and physical fitness) [23].

Mukherjee et al. [74] found that time-domain measures of HRV (i.e. mRR, SDNN, and RMSSD) were the most sensitive to changes in mental task difficulty, with mostly medium to large effect sizes between an "easy" and a "hard" trial of visual working memory tasks. This is consistent with our findings as well as the literature reporting that vm-HRV is sensitive to neurocognitive demands (e.g., difficulty, complexity, duration) related to cognitive and mental effort in older adults [28–31]. Our findings are also consistent with the conceptual framework of Silvestrini et al. [81] and the *"vagal tank theory"* [27] suggesting that vm-HRV may indeed be a valid biomarker of ITL during (exergame-based) simultaneous motor-cognitive training. However, the SEMs were often very large, which hampers the detection of changes over time [66]. Because it is commonly accepted that the SEM is a fixed characteristic of any measure, regardless of the sample of study participants under investigation [66], this indicates insufficient precision of the individual measurements, which currently limits the applicability of vm-HRV to monitor ITL when measured with portable HR monitors (e.g. Polar H10).

Implications for research

Although there is consistent evidence that HRV measurements obtained from the measurement of inter-beat-intervals through one-lead ECG via portable HR monitors shows a small amount of error compared to HRV derived from multi-lead ECG recordings [36, 52], further research is required to investigate the test-retest reliability of vm-HRV during exergame-based training and motor-cognitive training in general. In particular, future research should further investigate the reliability and validity of vm-HRV during exergame-based training and motor-cognitive training in general with a specific focus on comparing different measurement methodologies (e.g., measurement durations, technologies (i.e., measurement of inter-beat-intervals through one-lead ECG via portable heart rate monitor compared to multi-lead ECG recordings as well as different recording devices) as well as different analysis methodologies (e.g., beat correction and noise handling algorithms, or computation methods to calculated vm-HRV parameters), particularly focusing on ultra-short-term HRV measurements. Additionally, future research should more systemically evaluate the reliability and validity of vm-HRV under different exercise conditions (e.g., different levels of physical and/or cognitive task demands as well as targeted neurocognitive domains (e.g., as defined in [76] in line with the Diagnostic and Statistical Manual of Mental Disorders 5th Edition (DSM-5) [77]) and further examine the validity and potential implications of using vm-HRV as a biomarker of ITL during exergame-based training or motor-cognitive training in general. In particular, it should be investigated whether training that is prescribed and monitored according to real-time monitoring of ITL according to physiological parameters is superior to other markers for ITL and monitoring strategies, such as HRR, game metrics performance progression analysis or subjective ratings of ITL. These investigations have the potential to advance the utilization of vm-HRV in monitoring ITL during motor-cognitive training, and thereby pave the way to optimize individualized training prescription, reduce variability in training responses, and improve our understanding of the dose–response relationship between exercise and cognitive functioning [16].

Limitations

The study has some key limitations that are worth mentioning. First, the measurement of vm-HRV was done using a one-lead ECG via portable HR monitor, and multi-lead ECG data was not collected in parallel to assess the agreement of the outcome measures with ECG data, although multi-lead ECG is considered the gold standard for measuring HRV [51]. This approach was chosen due to consistent evidence demonstrating a small amount of absolute error in HRV measurements obtained from the measurement of inter-beat-intervals through one-lead ECG via portable HR monitors when compared to multi-lead ECG recordings [36, 52]. Additionally, portable HR monitors (e.g., chest belts) are widely spread and have good ease of use for monitoring ITL during everyday training. Nevertheless, the use of data from multi-lead ECG recordings may provide more accurate measurements of vm-HRV compared to one-lead ECG via portable HR monitor due to a reduction of movement artifacts and the measurement of raw ECG signal instead of solely the inter-beat-intervals [40]. Second, despite designing and pilot-testing the study protocol to mitigate learning effects (i.e., by (1) the inclusion of a standardized familiarization session; (2) the commencement of each block with a trial with adaptive task demands before the three standardized levels of external task demands that were evaluated, (3) the randomization of the order of all games, as well as (4) the randomization of the three standardized levels of external task demands within each game), the data revealed the occurrence of some learning effects, as there were main effects of time in the perceived task load with a decrease in the perceived task load from the test to retest measurement in some exergaming-conditions. Third, while the study followed recommendations to minimize the influence of transient confounding effects [40], it was impossible to check whether the participants adhered to these instructions.

This may have led to increased inter- and intra-individual variability of the vm-HRV measurements. Fourth, while we ensured that HR_{rest} remained within ± 5 bpm throughout the experimental session before starting a new exergame, we failed to do the same for vm-HRV. However, it has been reported that acute exercise effects post-exercise vm-HRV, which may be influenced by exercise intensity and/or duration [32, 82]. To mitigate the likelihood of any consequently biased outcomes, all standardized levels of external task demands as well as the five exergames were applied in randomized order. Finally, in the second experimental phase, vm-HRV values did not always reach a steady state for the last 60 s that were analyzed. This likely explains the mixed findings for test-retest reliability of vm-HRV during exergaming, as discussed in section 'Test-retest reliability of vm-HRV' and warrants future research to determine the minimum timeframe required to achieve steady state vm-HRV during exergaming in dependence on the physical and cognitive task demands.

Conclusion

Only the vm-HRV parameters mRR and PNS-Index demonstrated reliable measurement and served as valid biomarkers for quantifying ITL during exergame-based motor-cognitive training at a group level. Nonetheless, the presence of large SEMs hampers the detection of individual changes over time and suggests insufficient precision of these measurements at the individual level. These findings emphasize the potential and current limitations of vm-HRV as a biomarker for monitoring ITL during exergame-based training or motor-cognitive training in general. Future research should further investigate the reliability and validity of vm-HRV with a specific focus on comparing different measurement methodologies and exercise conditions. This should include, but not be limited to, different measurement durations, measurement technologies, and analysis methodologies, as well as varying physical and cognitive tasks and task demands (more details see "Implications for research" section). Additionally, the potential implications of using vm-HRV as a biomarker of ITL during exergame-based training or motor-cognitive training in general should be examined. In particular, it should be investigated whether training prescribed and monitored according to real-time monitoring of ITL according to physiological parameters is superior to other markers for ITL and monitoring strategies, such as HRR, game metrics performance progression analysis, or subjective ratings of ITL. These investigations have the potential to advance the utilization of vm-HRV in monitoring ITL during motor-cognitive training, thereby paving the way to optimize individualized training prescriptions, reduce variability in training responses, and improve our understanding of the dose–response relationship between exercise and cognitive functioning [16].

Abbreviations

ANOVA	Analysis of variance
CAN	Central autonomic network
ECG	Electrocardiogram
GCP	Good clinical practice
GRRAS	Guidelines for Reporting Reliability and Agreement Studies
HF	Absolute power of the high-frequency (0.15 – 0.4 Hz)
HF_{nu}	Relative power of HF (in normal units; HF [n.u.] = HF [ms^2] / (total power [ms^2] – very low frequency (0.00 – 0.04 Hz [ms^2]))
HOA	Healthy Older Adults
HR	Heart Rate
$\%HR_{max}$	Percentage of individual maximal heart rate
HR_{max}	Maximal heart rate
HRR	Heart rate reserve
HR_{rest}	Resting-state measurement of heart rate
HR_{target}	Target heart rate
ICC	Intraclass Correlation Coefficient
ITL	Internal training load
mRR	Mean R-R time interval
PNS-Index	Parasympathetic nervous system tone index
Qmci	Quick Mild Cognitive Impairment Screen
RMSSD	Root mean square of successive RR interval differences
RTLX	Raw Task Load Index
SD1	Poincaré plot standard deviation perpendicular to the line of identity
SDD	Smallest detectable difference
SDD%	Mean-normalized smallest detectable difference
SEM	Standard error of measurement
SEM%	Mean-normalized standard error of measurement
TLX	Task Load Index
vm-HRV	Vagally-mediated Heart Rate Variability

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13102-024-00929-y.

Supplementary Material 1.

Acknowledgements

The authors would like to thank all recruitment partners and all participants in this study for their participation and valuable contribution to this project. Additionally, the authors would like to thank André Groux, Karishma Thekkanath and Robin Mozolowski for their support in data collection.

Authors' contributions

PM conceptualized the study and was responsible for participant recruitment, data collection, statistical analysis, and writing the manuscript under the supervision of EdB. Both authors contributed to the revisions of the manuscript. Both authors read and approved the submitted version of the manuscript.

Funding

Open access funding provided by Swiss Federal Institute of Technology Zurich The authors received no specific funding for this work.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the Zenodo repository, https://doi.org/10.5281/zenodo.7824568.

Declarations

Ethics approval and consent to participate

All study procedures were carried out in accordance with the Declaration of Helsinki. The study protocol was approved by the ETH Zurich Ethics Committee (EK-2020-N-158).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Motor Control and Learning Group, Institute of Human Movement Sciences and Sport, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland. [2]Department of Health, OST - Eastern Swiss University of Applied Sciences, St. Gallen, Switzerland. [3]Division of Physiotherapy, Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Stockholm, Sweden.

Received: 18 June 2023 Accepted: 14 June 2024
Published online: 27 June 2024

References

1. World Health Organization. Global status report on the public health response to dementia; ISBN: 978-92-4-003324-5. 2021.
2. Veronese N, et al. Physical activity and exercise for the prevention and management of mild cognitive impairment and dementia: a collaborative international guideline. Eur Geriatr Med. 2023;14(5):925–52. https://doi.org/10.1007/s41999-023-00858-y.
3. Witherspoon L. ACSM information on exergaming. American College of Sports Medicine; 2013. p. 1. https://healthysd.gov/wp-content/uploads/2014/11/exergaming.pdf.
4. Stojan R, Voelcker-Rehage C. A systematic review on the cognitive benefits and neurophysiological correlates of exergaming in healthy older adults. J Clin Med. 2019;8(5):734. https://doi.org/10.3390/jcm8050734.
5. Temprado J-J. Can exergames be improved to better enhance behavioral adaptability in older adults? An ecological dynamics perspective. Front Aging Neurosci. 2021;13:670166. https://doi.org/10.3389/fnagi.2021.670166.
6. Torre MM, Temprado J-J. A review of combined training studies in older adults according to a new categorization of conventional interventions. Front Aging Neurosci. 2022;13:808539. https://doi.org/10.3389/fnagi.2021.808539.
7. Sokolov AA, et al. Serious video games and virtual reality for prevention and neurorehabilitation of cognitive decline because of aging and neurodegeneration. Curr Opin Neurol. 2020;33(2):239–48. https://doi.org/10.1097/WCO.0000000000000791.
8. Mishra J, et al. Video games for neuro-cognitive optimization. Neuron. 2016;90(2):214–8. https://doi.org/10.1016/j.neuron.2016.04.010.
9. Debettencourt MT, et al. Closed-loop training of attention with real-time brain imaging. Nat Neurosci. 2015;18(3):470–165. https://doi.org/10.1038/nn.3940.
10. Manser P, Herold F, de Bruin ED. Components of effective exergame-based training to improve cognitive functioning in middle-aged to older adults - a systematic review and meta-analysis. 2024. https://doi.org/10.1016/j.arr.2024.102385.
11. Torre MM, Temprado J-J. Effects of exergames on brain and cognition in older adults: a review based on a new categorization of combined training intervention. Front Aging Neurosci. 2022;14:859715. https://doi.org/10.3389/fnagi.2022.859715.
12. Foster C, et al. Monitoring training loads: the past, the present, and the future. Int J Sports Physiol Perform. 2017;12(s2):S2-2-S2-8. https://doi.org/10.1123/IJSPP.2016-0388.
13. Perrey S. Training monitoring in sports: it is time to embrace cognitive demand. Sports. 2022. https://doi.org/10.3390/sports10040056.
14. Herold F, et al. A discussion on different approaches for prescribing physical interventions – four roads lead to Rome, but which one should we choose? J Pers Med. 2020. https://doi.org/10.3390/jpm10030055.
15. Herold F, et al. New directions in exercise prescription: is there a role for brain-derived parameters obtained by functional near-infrared spectroscopy? Brain Sci. 2020. https://doi.org/10.3390/brainsci10060342.
16. Herold F, et al. Dose-response matters! – a perspective on the exercise prescription in exercise-cognition research. Front Psychol. 2019;10:2338. https://doi.org/10.3389/fpsyg.2019.02338.
17. Impellizzeri FM, et al. Internal and external training load: 15 years on. Int J Sports Physiol Perform. 2019;14(2):270–3. https://doi.org/10.1123/ijspp.2018-0935.
18. Netz Y. Is there a preferred mode of exercise for cognition enhancement in older age?—a narrative review. Front Med. 2019;6:57–57. https://doi.org/10.3389/fmed.2019.00057.
19. Garber CE, et al. American College of Sports Medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise. Med Sci Sports Exerc. 2011;43(7):1334–59. https://doi.org/10.1249/MSS.0b013e318213fefb.
20. Skulmowski A. Guidelines for choosing cognitive load measures in perceptually rich environments. Mind Brain Educ. 2022. https://doi.org/10.1111/mbe.12342.
21. Ayres P, et al. The validity of physiological measures to identify differences in intrinsic cognitive load. Front Psychol. 2021;12:702538. https://doi.org/10.3389/fpsyg.2021.702538.
22. Paas F, et al. Cognitive load measurement as a means to advance cognitive load theory. Educ Psychol. 2003;38(1):63–71. https://doi.org/10.1207/S15326985ep3801_8.
23. Manser P, et al. Can reactivity of heart rate variability be a potential biomarker and monitoring tool to promote healthy aging? A systematic review with meta-analyses. Front Physiol. 2021;12(1133):686129. https://doi.org/10.3389/fphys.2021.686129.
24. Thayer JF, Lane RD. A model of neurovisceral integration in emotion regulation and dysregulation. J Affect Disord. 2000;61(3):201–16. https://doi.org/10.1016/s0165-0327(00)00338-4.
25. Smith R, et al. The hierarchical basis of neurovisceral integration. Neurosci Biobehav Rev. 2017;75:274–96. https://doi.org/10.1016/j.neubiorev.2017.02.003.
26. Thayer JF. Heart rate variability: a neurovisceral integration model. In: RS Larry, editor. Encyclopedia of neuroscience. 2009. p. 1041–1047. https://doi.org/10.1016/B978-008045046-9.01991-4.
27. Laborde S, et al. Vagal tank theory: the three Rs of cardiac vagal control functioning - resting, reactivity, and recovery. Front Neurosci. 2018;12:458–458. https://doi.org/10.3389/fnins.2018.00458.
28. Hughes AM, et al. Cardiac measures of cognitive workload: a meta-analysis. Hum Factors. 2019;61(3):393–414. https://doi.org/10.1177/0018720819830553.
29. Ranchet M, et al. Cognitive workload across the spectrum of cognitive impairments: a systematic review of physiological measures. Neurosci Biobehav Rev. 2017;80:516–37. https://doi.org/10.1016/j.neubiorev.2017.07.001.
30. Castaldo R, et al. Acute mental stress assessment via short term HRV analysis in healthy adults: a systematic review with meta-analysis. Biomed Signal Process Control. 2015;18:370–7. https://doi.org/10.1016/j.bspc.2015.02.012.
31. Kim HG, et al. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig. 2018;15(3):235–45. https://doi.org/10.30773/pi.2017.08.17.
32. Michael S, et al. Cardiac autonomic responses during exercise and post-exercise recovery using heart rate variability and systolic time intervals-a review. Front Physiol. 2017;8:301. https://doi.org/10.3389/fphys.2017.00301.
33. Dong J-G. The role of heart rate variability in sports physiology. Exp Ther Med. 2016;11(5):1531–6. https://doi.org/10.3892/etm.2016.3104.
34. Gronwald T, Hoos O. Correlation properties of heart rate variability during endurance exercise: a systematic review. Ann Noninvasive Electrocardiol. 2019;25(1):e12697. https://doi.org/10.1111/anec.12697.
35. Georgiou K, et al. Can wearable devices accurately measure heart rate variability? A systematic review. Folia Med (Plovdiv). 2018;60(1):7–20. https://doi.org/10.2478/folmed-2018-0012.

36. Board L, et al. Validity of telemetric-derived measures of heart rate variability: a systematic review. J Exerc Physiol. 2016;19:64–84.
37. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Front Public Health. 2017;5:258–258. https://doi.org/10.3389/fpubh.2017.00258.
38. Kottner J, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. Int J Nurs Stud. 2011;48(6):661–71. https://doi.org/10.1016/j.ijnurstu.2011.01.016.
39. von Elm E, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453–7. https://doi.org/10.1016/S0140-6736(07)61602-X.
40. Laborde S, et al. Heart rate variability and cardiac vagal tone in psychophysiological research - recommendations for experiment planning, data analysis, and data reporting. Front Psychol. 2017;8:213. https://doi.org/10.3389/fpsyg.2017.00213.
41. Herold F, et al. Thinking while moving or moving while thinking - concepts of motor-cognitive training for cognitive performance enhancement. Front Aging Neurosci. 2018;10:228. https://doi.org/10.3389/fnagi.2018.00228.
42. Manser P, de Bruin ED. Making the best out of IT: design and development of exergames for older adults with mild neurocognitive disorder - a methodological paper. Front Aging Neurosci. 2021;13:734012. https://doi.org/10.3389/fnagi.2021.734012.
43. Dividat AG. Vimeo - Dividat AG. 2022. https://vimeo.com/dividat. Accessed 28 Feb 2022.
44. Karvonen J, Vuorimaa T. Heart rate and exercise intensity during sports activities. Practical application. Sports Med. 1988;5(5):303–11. https://doi.org/10.2165/00007256-198805050-00002.
45. Karvonen MJ, et al. The effects of training on heart rate; a longitudinal study. Ann Med Exp Biol Fenn. 1957;35(3):307–15.
46. Manser P, de Bruin ED. Diagnostic accuracy, reliability, and construct validity of the German Quick Mild Cognitive Impairment Screen [submitted for publication, under review]. 2024. https://doi.org/10.13140/RG.2.2.27316.63369.
47. O'Caoimh R. The Quick Mild Cognitive Impairment (Qmci) Screen: developing a new screening test for mild cognitive impairment and dementia. University College Cork; 2015. https://hdl.handle.net/10468/2170.
48. O'Caoimh R, Molloy DW. The Quick Mild Cognitive Impairment Screen (Qmci). In: Cognitive screening instruments. 2017. p. 255–272.
49. O'Caoimh R, et al. The Quick Mild Cognitive Impairment Screen correlated with the standardized Alzheimer's disease assessment scale–cognitive section in clinical trials. J Clin Epidemiol. 2014;67(1):87–92. https://doi.org/10.1016/j.jclinepi.2013.07.009.
50. Glynn K, et al. Is the Quick Mild Cognitive Impairment Screen (QMCI) more accurate at detecting mild cognitive impairment than existing short cognitive screening tests? A systematic review of the current literature. Int J Geriatr Psychiatry. 2019;34(12):1739–46. https://doi.org/10.1002/gps.5201.
51. Mosley E, Laborde S. A scoping review of heart rate variability in sport and exercise psychology. Int Rev Sport Exerc Psychol. 2022:1–75. https://doi.org/10.1080/1750984X.2022.2092884.
52. Dobbs WC, et al. The accuracy of acquiring heart rate variability from portable devices: a systematic review and meta-analysis. Sports Med. 2019;49(3):417–35. https://doi.org/10.1007/s40279-019-01061-5.
53. Malik M. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. Circulation. 1996;93:1043–65.
54. Williams DP, et al. Two-week test-retest reliability of the Polar® RS800CX™ to record heart rate variability. Clin Physiol Funct Imaging. 2017;37(6):776–81. https://doi.org/10.1111/cpf.12321.
55. Lipponen JA, Tarvainen MP. A robust algorithm for heart rate variability time series artefact correction using novel beat classification. J Med Eng Technol. 2019;43(3):173–81. https://doi.org/10.1080/03091902.2019.1640306.
56. Niskanen J-P, et al. Software for advanced HRV analysis. Comput Methods Programs Biomed. 2004;76(1):73–81. https://doi.org/10.1016/j.cmpb.2004.03.004.
57. Tarvainen MP, et al. Kubios HRV - heart rate variability analysis software. Comput Methods Programs Biomed. 2014;113(1):210–20. https://doi.org/10.1016/j.cmpb.2013.07.024.
58. Tarvainen MP, et al. An advanced detrending method with application to HRV analysis. IEEE Trans Biomed Eng. 2002;49(2):172–5. https://doi.org/10.1109/10.979357.
59. Ernst G. Heart-rate variability-more than heart beats? Front Public Health. 2017;5:240. https://doi.org/10.3389/fpubh.2017.00240.
60. Tarvainen MP, Niskanen J-P, Ranta-aho PO. Kubios HRV (ver. 3.4) user's guide. 2018.
61. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock PA, Meshkati N, editors. Human mental workload. Advances in psychology. North-Holland; 1988. p. 139–83. https://doi.org/10.1016/S0166-4115(08)62386-9.
62. Hart SG. Nasa-Task Load Index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet. 2016;50(9):904–8. https://doi.org/10.1177/154193120605000909.
63. Thompson CB. Descriptive data analysis. Air Med J. 2009;28(2):56–9. https://doi.org/10.1016/j.amj.2008.12.001.
64. Mishra P, et al. Descriptive statistics and normality tests for statistical data. Ann Card Anaesth. 2019;22(1):67–72. https://doi.org/10.4103/aca.ACA_157_18.
65. Field A, et al. Discovering statistics using R. Sage publications; 2012. https://us.sagepub.com/en-us/nam/discovering-statistics-using-r/book236067.
66. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231–40. https://doi.org/10.1519/15184.1.
67. Noguchi K, et al. nparLD: an R software package for the nonparametric analysis of longitudinal data in factorial experiments. J Stat Softw. 2012;50(12):1–23. https://doi.org/10.18637/jss.v050.i12.
68. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012.
69. Shrout PE, Fleiss JL. Intraclass correlations - uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–8. https://doi.org/10.1037/0033-2909.86.2.420.
70. Cohen J. Statistical power analysis for the behavioral sciences; ISBN: 1134742703. Routledge; 1988. https://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf.
71. Rosenthal R. Meta-analytic procedures for social research. Thousand Oaks: SAGE Publications, Inc; 1991.
72. Borg DN, et al. Calculating sample size for reliability studies. PM R. 2022;14(8):1018–25. https://doi.org/10.1002/pmrj.12850.
73. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med. 2002;21(9):1331–5. https://doi.org/10.1002/sim.1108.
74. Mukherjee S, et al. Sensitivity to mental effort and test–retest reliability of heart rate variability measures in healthy seniors. Clin Neurophysiol. 2011;122(10):2059–66. https://doi.org/10.1016/j.clinph.2011.02.032.
75. Guijt AM, et al. Test-retest reliability of heart rate variability and respiration rate at rest and during light physical activity in normal subjects. Arch Med Res. 2007;38(1):113–20. https://doi.org/10.1016/j.arcmed.2006.07.009.
76. Sachdev PS, et al. Classifying neurocognitive disorders: the DSM-5 approach. Nat Rev Neurol. 2014;10(11):634–42. https://doi.org/10.1038/nrneurol.2014.181.
77. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub; 2013. https://repository.poltekkes-kaltim.ac.id/657/1/Diagnostic%20and%20statistical%20manual%20of%20mental%20disorders%20_%20DSM-5%20(%20PDFDrive.com%20).pdf.
78. Nunan D, et al. A quantitative systematic review of normal values for short-term heart rate variability in healthy adults. Pacing Clin Electrophysiol. 2010;33(11):1407–17. https://doi.org/10.1111/j.1540-8159.2010.02841.x.
79. Hou C-J, et al. Analysis of heart rate variability in response to serious games in elderly people. Sensors. 2021;21(19):6549.
80. Hou C-J, et al. Analysis of heart rate variability and game performance in normal and cognitively impaired elderly subjects using serious games. Appl Sci. 2022;12(9):4164.
81. Silvestrini N. Psychological and neural mechanisms associated with effort-related cardiovascular reactivity and cognitive control: an integrative approach. Int J Psychophysiol. 2017;119:11–8. https://doi.org/10.1016/j.ijpsycho.2016.12.009.
82. Singh N, et al. Heart rate variability: an old metric with new meaning in the era of using mHealth technologies for health and exercise training guidance. Part two: prognosis and training. Arrhythm Electrophysiol Rev. 2018;7(4):247–55. https://doi.org/10.1542/aer.2018.30.2.

Publisher's Note