

Ger J Exerc Sport Res  
<https://doi.org/10.1007/s12662-020-00682-7>  
Eingegangen: 19. April 2020  
Angenommen: 6. Oktober 2020

© Der/die Autor(en) 2020



Matthias Wagner<sup>1</sup> · Florian Engel<sup>2</sup> · Kristina Klier<sup>3</sup> · Saskia Klughardt<sup>2</sup> · Franziska Wallner<sup>1</sup> · Alissa Wiczorek<sup>1</sup>

<sup>1</sup> Fakultät für Humanwissenschaften, Institut für Sportwissenschaft, Forschungs- und Lehrgebiet Sportpsychologie, Universität der Bundeswehr München, Neubiberg, Deutschland

<sup>2</sup> Fakultät für Humanwissenschaften, Institut für Sportwissenschaft, Forschungs- und Lehrgebiet Sportmethodik/Trainingswissenschaft, Universität der Bundeswehr München, Neubiberg, Deutschland

<sup>3</sup> Institut für Sportwissenschaft, Forschungs- und Lehrgebiet Gesundheit, Betriebliche Gesundheitsförderung und Prävention, Fakultät für Humanwissenschaften, Universität der Bundeswehr München, Neubiberg, Deutschland

# Zur Reliabilität von Wearable Devices am Beispiel einer Premium Multisport-Smartwatch

## Einleitung

### Hintergrund

Gesundheit, Aktivität und Fitness kommt gerade im Zuge des digitalen Wandels eine zentrale individuelle, aber auch gesamtgesellschaftliche Bedeutung zu. Neben den antizipierten Risiken der Digitalisierung, welche hier mit Inaktivitätsbedingten Fitness- und Gesundheitsdefiziten umrissen werden (vgl. Pratt et al., 2012), zeugen die stetig wachsenden Absatzzahlen von Wearable Devices (WD; u. a. Lutter, Meineke, & Tropf, 2017) sowie die Kombination ebd. mit entsprechend kompatiblen digitalen Diensten (u. a. Twitter; vgl. Chung, Skinner, Hasty, & Perrin, 2016) jedoch gleichsam vom Bestreben, mögliche Potenziale der Digitalisierung für die positive Entwicklung der Parameter Gesundheit, Aktivität und Fitness nutzbar zu machen. Unter WD werden im Folgenden alle Formen technischer Sensoren verstanden, die am Körper oder der Kleidung getragen werden und die Erfassung gesundheits-, aktivitäts- und fitnessbezogener Daten ermöglichen (vgl. Peake, Kerr, & Sullivan, 2018; Patel, Park, Bonato, Chan, & Rodgers, 2012).

### Forschungsrahmen

Der grundlegende Zusammenhang zwischen Gesundheit, Aktivität und Fitness

kann über ein systemisches Anforderungs-Ressourcen-Modell der Gesundheit in Anlehnung an Becker (1992) unter Einbeziehung der im sportwissenschaftlichen Themenfeld Bewegung und Gesundheit entstandenen Modellerweiterungen (u. a. Becker, Bös, & Woll, 1994) begründet werden. Gesundheitliche Ressourcen werden in ebd. Modellen u. a. in der körperlichen Leistungsfähigkeit, gesundheitliche Anforderungen demgegenüber u. a. im Faktor Übergewicht gesehen. Potenziell interventionswirksame Determinanten zur Aufrechterhaltung eines gesundheitsförderlichen Maßes an körperlich-sportlicher Aktivität finden sich in der sozial-kognitiv ausgerichteten Physical Activity Maintenance Theory (PAMT; Nigg, Borrelli, Maddock, & Dishman, 2008). Der PAMT-Modellkomponente Motivation kann in einem durch WD gerahmten Aktivitätskontext eine besondere Bedeutung beigemessen werden, da die Nutzung von WD, gerade bei Personen, die nach einem sicheren Bezugsrahmen für ihr Handeln streben und geringe Aussicht auf Erfolg in sich verspüren (in der Regel Neueinsteiger), motivierend wirken kann (vgl. Attig & Franke, 2019). Dem aus der Literatur bekannten Phänomen des Drop-outs innerhalb der ersten sechs Monaten nach Aufnahme einer neuen Aktivität (vgl. Dishman & Buckworth, 1996) kann von technischer Seite und in Anlehnung an die Self Determination Theory (Deci &

Ryan, 1985) potenziell dann entgegengewirkt werden, wenn WD dem Nutzer bei seinen Aktivitäten dauerhafte Autonomie, Kompetenzerleben und soziale Einbindung ermöglichen. Diese Potenziale scheinen insofern gegeben, da WD die Formulierung individueller Fitness-, Aktivitäts- und Gesundheitsziele und somit (scheinbar) autonomes Handeln ermöglichen, das Kompetenzerleben durch die Spiegelung der eigenen Gesundheits-, Aktivitäts- und Fitnessentwicklung in Form von Challenge- und Zielerreichungsszenarien befördern und über das Teilen der eigenen Gesundheits-, Aktivitäts- und Fitnessdaten in virtuellen Communities eine soziale Einbindung versprechen. Gleichwohl ist an dieser Stelle festzuhalten, dass sich die empirische Befundlage zu den motivationspsychologischen Aspekten der Nutzung von WD (u. a. Kerner & Goodyear, 2017; Busch, Utesch, Bürkner, & Strauss, 2020) bislang noch wenig einheitlich gestaltet. Gerade das Kompetenzerleben ist bei der Aufrechterhaltung der Motivation zur körperlichen Aktivität insofern bedeutsam, da in Anlehnung an das Expectancy-Value Model of Achievement (Eccles & Wigfield, 2002) erst die Erfahrung der eigenen Kompetenz, interne Erfolgsattributionen, positive emotionale Zustände sowie positive Effekte auf das physische Selbstkonzept und die sport- und bewegungsbezogene Selbstwirksamkeit und damit den Prozess der

Internalisierung hin zu einer intrinsischen Handlungsmotivation ermöglicht. Das Erleben der eigenen Kompetenz kann somit initial zur Aufrechterhaltung ressourcenstärkender bzw. anforderungsmindernder Aktivitäten (auch und gerade) im digitalen Zeitalter beitragen. Kompetenzerleben setzt jedoch die Möglichkeit zur systematischen Evaluation des eigenen Trainingsprozesses voraus (u. a. Gerber, 2014). Für einen durch WD gerahmten Aktivitätskontext und aus Perspektive der klassischen Testtheorie bedeutet dies, dass Kompetenzerleben nur dann entstehen kann, wenn WD auch möglichst reliabel und valide den jeweils aktuellen psychophysischen Zustand des Nutzers spiegeln; an dieser Stelle setzt die vorliegende Arbeit an.

### Forschungsfrage und Forschungshypothese

Nachhaltige gesundheitsförderliche Effekte körperlich-sportlicher Aktivitäten setzen einen strukturierten und systematischen Trainingsprozess voraus (Brehm et al., 2013); im Sinne einer diagnostisch-evaluativ akzentuierten trainingswissenschaftlichen Analogie sei hier auf das kybernetisch begründete Regelkreismodell der Trainingssteuerung (u. a. Hohmann, Lames, & Letzelter, 2014) verwiesen. In der Elaboration des erweiterten Themenfeldes Bewegung, Gesundheit und Digitalisierung anhand von WD kommt der funktionshöchsten Klasse der Puls- bzw. Sportuhren demnach insofern eine besondere Bedeutung zu, da diese über die Dokumentation der Bewegungs- und Sportaktivitäten hinaus gleichsam die Möglichkeit zur Analyse und Steuerung ebensolcher Trainingsprozesse versprechen (vgl. Aroganam, Manivannan, & Harrison, 2019). Aus motivationspsychologischer und trainingswissenschaftlicher Perspektive erscheint es dabei, wie oben deduziert, angezeigt, die Testgüte von WD am Beispiel eines marktrelevanten Gerätes zu prüfen. In der Sondierungsphase der zu diesem Zwecke konzipierten, eigenen empirischen Untersuchung wurde die Marke Garmin (Garmin Deutschland GmbH, Garching) als etablierter Key-

Player im Fitness-Segment (vgl. Statista, 2020, S. 31) sowie deren Premium Multisport-Smartwatch Garmin fenix® 5 (GF5; Modelljahr 2017) als im Unternehmensportfolio besonders absatzstarkes Produkt (vgl. Statista, 2020, S. 32) identifiziert. Die GF5 inkludiert eine Vielzahl unterschiedlicher Funktionen zur Messung gesundheits-, aktivitäts- und fitnessbezogener Daten, wobei in den GF5-Parametern Stress – als psychophysische Reaktion auf digitale Anforderungen – sowie Kalorienverbrauch und maximale Sauerstoffaufnahme (VO<sub>2</sub>max) – als primäre Steuerungsindikatoren zur Gewichtsreduktion und Leistungsentwicklung – die größte Aussagekraft im Hinblick auf die Beurteilung der hier interessierenden Untersuchungsbereiche Gesundheit, Aktivität und Fitness vermutet werden kann. Zur Beantwortung der Forschungsfrage nach der Testgüte der GF5 muss festgehalten werden, dass ebd. zwar zuletzt methodisch angemessen im Hinblick auf die Validität der Messung von Herzfrequenz und Energieverbrauch überprüft wurde (vgl. Düking et al., 2020), Befunde zur Reliabilität und dabei explizit zum Reliabilitätsaspekt der Reproduzierbarkeit (*engl.*: „reproducibility“) von referenzmethodenbasierten Messergebnissen seitens der GF5 bislang jedoch nicht vorliegen. Die Sichtung der Whitepapers zur Überprüfung der in die GF5 integrierten Algorithmen zur Bestimmung von Stress (vgl. Firstbeat Technologies Ltd., 2014), Kalorienverbrauch (vgl. Firstbeat Technologies Ltd., 2012) und VO<sub>2</sub>max (vgl. Firstbeat Technologies Ltd., 2017) führten zu der globalen Forschungshypothese einer grundlegenden Übereinstimmung (*engl.*: „concordance“) zwischen den Messergebnissen der GF5 und den Messergebnissen etablierter Referenzmethoden.

### Methodische Vorbemerkung

An dieser Stelle möchten die Autorinnen und Autoren dieses Beitrages darauf hinweisen, dass der Vergleich mit einer Referenzmethode oftmals reflexartig mit dem Ansinnen nach konvergenter Validierung im Sinne der von Campbell und Fiske (1959) vorgeschlagenen Mul-

titrait-Multimethod-Analyse assoziiert wird, dieser methodische Zugang jedoch auch und zunächst einmal die Reproduzierbarkeit von referenzmethodenbasierten Messergebnissen und somit die hier adressierte Überprüfung der Übereinstimmung (*engl.*: „agreement“, syn.: „concordance“) zwischen unterschiedlichen Methoden (*engl.*: „between-method agreement“) im Sinne der Reliabilität begründet und ermöglicht. Anstelle der im Kontext der Monotrait-Heteromethodbasierten Konvergenzüberprüfung üblichen (linearen) Zusammenhangsmaße sind im Falle der Konkordanzüberprüfung entsprechend davon abweichende Übereinstimmungsmaße zu diskutieren (vgl. Liu et al., 2016). Eine hohe Korrelation zwischen unterschiedlichen Methoden (u. a. Düking et al., 2020) kann folgerichtig zwar als Hinweis auf konvergente Validität interpretiert werden; zu berücksichtigen ist an dieser Stelle jedoch, dass der implizite Schluss auf ein gleichsam reliables Instrument im Hinblick auf den Reliabilitätsaspekt der Konkordanz nicht gegeben sein muss („Two sets of observations, which are highly correlated, may have poor agreement“; Ranganathan, Pramesh, & Aggarwal, 2017, S. 187), weshalb eine dahingehend differenzierte Reliabilitätsanalyse für den Gesamtprozess der „Validierung“ unabdingbar ist.

Konkordanzbegründete Reliabilitätsbeurteilungen sind in den medizinisch-biowissenschaftlichen Fachdisziplinen fest etabliert und detailreich dokumentiert (im Überblick Watson & Petrie, 2010), werden in dem hier definierten gesundheits- und verhaltenswissenschaftlichen Kontext – trotz eindeutiger Differenzierungspotenziale gegenüber „klassischen“ Beurteilungsbegründungen – jedoch bislang kaum berücksichtigt (vgl. Barchard, 2012). Insofern wollen die Autorinnen und Autoren dieses Beitrages den von ihnen gewählten methodisch-analytischen Ansatz auch als Anreiz verstanden wissen, den oftmals in disziplinspezifischen Grenzen definierten Reliabilitätsbegriff und das damit verbundene Standardrepertoire an methodischen Zugängen (u. a. Düking, Fuss, Holmberg, & Sperlich, 2018) im interdisziplinären Diskurs zu erweitern.

M. Wagner · F. Engel · K. Klier · S. Klughardt · F. Wallner · A. Wiczorek

## Zur Reliabilität von Wearable Devices am Beispiel einer Premium Multisport-Smartwatch

### Zusammenfassung

Wearable Devices versprechen durch ihre motivierende Wirkung einen wichtigen Beitrag zur Bindung des Individuums an körperlich-sportliche Aktivitäten und somit zum Aufbau und Erhalt von Gesundheit und Leistungsfähigkeit in Zeiten des digitalen gesellschaftlichen Wandels. Übergeordnetes Ziel der vorliegenden Untersuchung war die Beurteilung der Testgüte von Wearable Devices anhand eines marktrelevanten Gerätes, der Garmin fenix® 5. Als Forschungsdesiderat wurde der Reliabilitätsaspekt der Methodenkonkordanz identifiziert. Zur Überprüfung der Methodenkonkordanz wurden das Stresslevel bei kognitiver Stressinduktion, der Kalorienverbrauch bei moderater Ausdauerlaufbelastung sowie die maximale

Sauerstoffaufnahme bei Laufausbelastung von 30 männlichen Probanden (Alter:  $23,13 \pm 2,5$  Jahre; BMI:  $24,95 \pm 2,45$  kg/m<sup>2</sup>) mit der Garmin fenix® 5 bestimmt und die Ergebnisse mit denen im Feld gängigen Referenzmethoden Elektrokardiographie, Indirekte Kalorimetrie bzw. Spiroergometrie verglichen. Zur rechnerischen Überprüfung der Methodenkonkordanz diente Lin's Konkordanzkorrelationskoeffizient ( $CCC_{Lin}$ ). Die Ergebnisse zeigen eine hohe Präzision der Garmin fenix® 5 im Vergleich mit der Referenzmethode Elektrokardiographie hinsichtlich der Messung des notwendigerweise z-standardisierten Stressparameters ( $p = 0,89$ ) sowie eine gerade mittlere exakte intraindividuelle Konkordanz mit der

Referenzmethode Indirekte Kalorimetrie bzw. Spiroergometrie hinsichtlich der Messung des Parameters Kalorienverbrauch ( $CCC_{Lin} = 0,43$  [ $p = 0,52$ ,  $C_b = 0,82$ ]) bzw. maximale Sauerstoffaufnahme ( $CCC_{Lin} = 0,50$  [ $p = 0,77$ ,  $C_b = 0,66$ ]). Die Garmin fenix® 5 kann somit zumindest bei erstmaliger Verwendung nicht als hinreichend konkordante Alternative zu den gängigen aktivitäts- und leistungsbezogenen Referenzmethoden empfohlen werden.

### Schlüsselwörter

Wearable Computing · Wearables · Fitnessstracker · Methodenkonkordanz · Goldstandard

## On the reliability of wearable devices using the example of a premium multisport smartwatch

### Abstract

In view of their motivating effect, wearable devices promise to provide an important contribution to individual's maintenance of physical activity and thus, to the strengthening of health- and performance-relevant parameters in times of digital change. The overall aim of the present study was to evaluate the psychometric properties of wearable devices using a strong selling product of one of the markets key-players, the Garmin fenix® 5. The reliability aspect of method concordance was identified as a research desideratum. To evaluate method concordance, stress level under cognitive stress exposure, energy expenditure during moderate running as well as maximum oxygen uptake during

running of 30 male participants (age:  $23.13 \pm 2.5$  years; BMI:  $24.95 \pm 2.45$  kg/m<sup>2</sup>) as measured by the Garmin fenix® 5 was compared with these parameters on the basis of the commonly accepted reference methods electrocardiography, indirect calorimetry and spiroergometry using Lin's Concordance Correlation Coefficient ( $CCC_{Lin}$ ). Following our results, the Garmin fenix® 5 does provide a high precision when compared to the reference method electrocardiography and with regard to the measurement of the necessarily z-standardized stress parameter ( $p = 0.89$ ). Moreover, indications for just moderate exact intraindividual concordance between the Garmin fenix® 5

and its respective reference method indirect calorimetry and spiroergometry regarding the parameters energy expenditure ( $CCC_{Lin} = 0.43$  [ $p = 0.52$ ,  $C_b = 0.82$ ]) and maximum oxygen uptake ( $CCC_{Lin} = 0.50$  [ $p = 0.77$ ,  $C_b = 0.66$ ]) were found. Therefore, the Garmin fenix® 5 cannot be recommended as a sufficiently concordant alternative to the established activity- and performance-related reference methods, at least when used for the first time.

### Keywords

Wearable computing · Wearables · Fitnessstracker · Method concordance · Goldstandard

## Methode

### Untersuchungsstichprobe

Zur Überprüfung der Reliabilität der GF5 hinsichtlich der Parameter Stress, Kalorienverbrauch und VO<sub>2</sub>max wurden männliche Probanden im „sich entwickelnden“ (engl.: „emerging adulthood“; 18 bis 25 Jahre; vgl. Lang, Martin, & Pinquart, 2012) und frühen (26 bis 40 Jahre; vgl. Lang et al., 2012) Erwachsenenalter im institutionellen Kontext der Autorinnen und Autoren

über einen Zeitraum von acht Wochen rekrutiert. Die Probanden mussten die Kompetenz zur Anwendung der deutschen Sprache in Wort und Schrift, ihre Rechtsfähigkeit, ihr Verständnis von Art, Umfang, Bedeutung und Risiko der Untersuchung sowie die Freiwilligkeit ihrer Teilnahme schriftlich bestätigen. Als Studienausschlusskriterien wurden das weibliche Geschlecht, ein während der Anamnese gemessener Blutdruck über 140/90 mm Hg (vgl. Banegas et al., 2018), die selbstberichtete Einnahme wahrnehmungsverändernder Substan-

zen sowie gleichsam selbstberichtete akute oder nicht vollständig ausgeheilte (Sport-)Verletzungen, Operationen oder schwerwiegende kardiovaskuläre, pulmonale, renale, neurologische, psychische bzw. degenerative Erkrankungen definiert.

### Untersuchungsablauf

Die Hauptuntersuchung untergliederte sich in drei aufeinanderfolgende Laborszenarien an Testtag 1 sowie ein Feldszenario an Testtag 2. Die beiden Test-

tage wurden im Abstand von minimal 24 Stunden (vgl. Anderson, Chisenall, Tolbert, Ruffner, Whitehead., & Conners, 2019) und maximal einer Woche realisiert. Die GF5 wurde von allen Probanden am linken Handgelenk getragen und vor jeder Testung individuell anhand der Parameter Alter, Größe, Gewicht und Geschlecht eingestellt.

Laborszenario (LS) 1 diente der Überprüfung der Reliabilität der GF5 hinsichtlich des Parameters Stress; zur Realisierung eines – dem Forschungsrahmen angemessenen – digitalen Stress-Szenarios absolvierten die Probanden eine modifizierte Version des Paced Auditory Serial Addition Task (PASAT) Test (PASAT-C; Lejuez, Kahler, & Brown, 2003). Beim PASAT-C handelt es sich um einen computergestützten Zahlenadditionstest, dessen stressierende Wirkung durch die mehrmalige Verkürzung der Latenzzeit bei der Ziffernpräsentation entsteht und anhand der Veränderung der Herzfrequenz (HF) während bzw. über den Unterschied der Stresseinschätzung vor und während der Exposition (vgl. Kolotylova et al., 2010) sowie über die Abnahme der Herzratenvariabilität (HRV) bei unterschiedlichen Expositionszeiten (vgl. Tanosoto et al., 2015) validiert wurde.

LS 2 diente der Überprüfung der Reliabilität der GF5 hinsichtlich des Parameters Kalorienverbrauch; hierzu absolvierten die Probanden ein moderates 20-minütiges Laufband-Belastungsprotokoll bei konstanter Laufgeschwindigkeit von 6 km/h und konstanter Steigung von 1 % zur Kompensation des fehlenden Luftwiderstandes (vgl. Meyer, 2007).

Der Übergang zwischen LS 2 und LS 3 wurde durch eine 2-minütige Laufbandaktivität bei 6 km/h Laufgeschwindigkeit sowie einer anschließenden Warm-up-Phase von einer Minute bei 10 km/h Laufgeschwindigkeit und 1 % Steigung gestaltet.

LS 3 diente der Überprüfung der Reliabilität der GF5 hinsichtlich des Parameters  $VO_2\max$ ; hierzu absolvierten die Probanden ein Rampenprotokoll bis zur subjektiven Erschöpfung (vgl. Hollmann & Strüder, 2009) auf dem Laufband, bei dem die Laufbandsteigung (Ausgangssteigung: 5 %) unter Konstanthaltung

der Laufgeschwindigkeit (10 km/h) ab der zweiten Minute kontinuierlich um 2,5 % pro Minute erhöht wurde (vgl. Kroidl, Schwarz, Lehnigk, & Fritsch, 2015). Zur objektiven Beurteilung der tatsächlichen individuellen maximalen Ausbelastung erfolgte eine Orientierung an den Richtlinien des American College of Sports Medicine (vgl. Riebe, Ehrman, Liguori, & Magal, 2018). Hiernach wurde eine Ausbelastung attestiert, wenn zwei der vier nachfolgenden Kriterien erfüllt waren: eine Plateaubildung der  $VO_2$  bei simultaner Leistungssteigerung im Echtzeit-Monitoring der Untersuchung (Levelling-off), eine respiratorische Gasaustauschrate von  $\geq 1,10$ , eine subjektive Erschöpfung nach Borg (1998) von  $\geq 17$  auf einer Skala von 6–20 sowie eine Verringerung der HF-Zunahme bei Annäherung an die individuelle Maximalleistung. Nach Abbruch des Laufband-Ausbelastungstests absolvierten die Probanden eine 1-minütige aktive Cool-down-Phase bei 4 km/h Laufgeschwindigkeit und einer Laufbandsteigung von 1 %.

Die beiden in LS 2 und LS 3 eingesetzten Protokolle wurden auf einem Pulsar<sup>®</sup>3p Laufband der Firma h/p/cosmos sports & medical GmbH (Nussdorf-Traunstein) realisiert.

Das an Testtag 2 konzipierte Feldszenario komplettierte die in LS 3 begonnene Überprüfung der Reliabilität der GF5 hinsichtlich des Parameters  $VO_2\max$ ; hierzu wurden die Probanden in Orientierung an Anderson et al. (2019) angewiesen, sich zunächst für zwei Runden bei individueller moderater Intensität auf einer genormten 400 m Tartanbahn einzulassen. In der anschließenden Testphase mussten die Probanden 10 min und 30 s bei individuell maximal möglicher Laufgeschwindigkeit auf der 400 m Tartanbahn absolvieren. Zwei abschließend im individuellen Gehtempo absolvierte Laufbahnstunden dienten der Erholung des Herz-Kreislauf-Systems.

## Messmethoden

Das GF5-Setup bestand aus der Handgelenkssmartwatch sowie dem zugehörigen HF-Brustgurt HRM-Run der Firma Garmin Deutschland GmbH (Garching).

Zur Überprüfung der Reliabilität der GF5 hinsichtlich des Parameters Stress in LS 1 diente die Referenzmethode der Elektrokardiographie (EKG; vgl. Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology, 1996). Das hierfür in LS 1 eingesetzte 12-Kanal EKG-Messsystem *custo cardio 300* der Firma custo med GmbH (Ottobrunn) ermöglichte eine beat-to-beat Aufzeichnung der HRV während der PASAT-C-Stressexposition.

Zur Überprüfung der Reliabilität der GF5 hinsichtlich des Parameters Kalorienverbrauch in LS 2 diente die Referenzmethode der indirekten Kalorimetrie (IK; vgl. Hills, Mokhtar, & Byrne, 2014). Das hierfür parallel zur GF5 in LS 2 eingesetzte, spiroergometrische Messsystem bestand aus einem Spirometer mit „Breath-by-breath“-Technologie (METALYZER<sup>®</sup> 3B, CORTEX Biophysik GmbH, Leipzig), einem BT Smart HF-Gurt (Polar H7) der Polar Electro GmbH Deutschland (Büttelborn), einer Atemmaske für Metalyzer 3B und MetaMax 3B mit Befestigungsnetz sowie einer multifunktionalen Anwendungssoftware (MetaSoft<sup>®</sup> Studio) der CORTEX Biophysik GmbH (Leipzig). Belege für die Messgenauigkeit der eingesetzten Hardware finden sich bei Meyer, Georg, Becker, und Kindermann (2001).

Zur Überprüfung der Reliabilität der GF5 hinsichtlich des Parameters  $VO_2\max$  diente die Referenzmethode der Spiroergometrie (SPE; vgl. Scharhag-Rosenberger, & Schommer, 2013). Das hierfür in LS 3 eingesetzte, spiroergometrische Gesamtsetup entsprach dem aus LS 2.

## Datenverarbeitung

Zur Messung des Parameters Stress nutzt die GF5 die über die integrierten Sensoren erfassten HRV-Informationen; diese werden im Rahmen eines 3-minütigen Quick-Level-Stress-Tests (QLST; vgl. Firstbeat Technologies Ltd., 2020) in einen Stressreaktionswert zwischen 1 (sehr geringe Belastung) und 100 (sehr hohe Belastung) überführt (vgl. Garmin Ltd., 2017). Die zur Überprüfung der Reliabilität der GF5 hinsichtlich des Parameters Stress in LS 1 über die Referenzmethode der Elektrokardiographie (EKG) erfassten HRV-Informationen werden im Rahmen eines 3-minütigen Quick-Level-Stress-Tests (QLST; vgl. Firstbeat Technologies Ltd., 2020) in einen Stressreaktionswert zwischen 1 (sehr geringe Belastung) und 100 (sehr hohe Belastung) überführt (vgl. Garmin Ltd., 2017). Die zur Überprüfung der Reliabilität der GF5 hinsichtlich des Parameters  $VO_2\max$  in LS 3 eingesetzte, spiroergometrische Gesamtsetup entsprach dem aus LS 2.

renzmethode EKG anhand der Analysesoftware Kubios (Mindfeld Biosystems Ltd., Gronau; vgl. Tarvainen, Niskanen, Lipponen, Ranta-Aho, & Karjalainen, 2014) ermittelten RR-Intervalle wurden im Rahmen einer Spektralanalyse mittels Fast-Fourier-Transformation in den Referenzparameter *High Frequency* (HF;  $\text{ms}^2$ ) überführt. HF findet bei HRV-Kurzzeitmessungen seine Anwendung, spiegelt Frequenzen in einer Range von 0,15–0,40 Hz wider und gilt bei niedriger Power als Indikator für Stress, Panik, Angst oder Sorgen (vgl. Shaffer & Ginsberg, 2017). Hierbei ist zu berücksichtigen, dass der in den ersten drei Minuten der Stressexposition über die GF5 gemessene QLST-Stressreaktionswert im Gegensatz zum, im gleichen Zeitintervall ermittelten, Referenzparameter HF mit steigender Stressreaktivität ansteigt. Aufgrund der unterschiedlichen Metriken wurden die QLST- bzw. HRV-Rohwerte mittels z-Transformation standardisiert und die z-standardisierten QLST-Werte invertiert.

Zur Messung des Parameters Kalorienverbrauch (kcal) nutzt die GF5 ein HF-basiertes neuronales Netzwerk (vgl. Firstbeat Technologies Ltd., 2012). Der zur Überprüfung der Reliabilität der GF5 hinsichtlich des Parameters Kalorienverbrauch in LS 2 über die Referenzmethode IK gemessene Referenzparameter Kalorienverbrauch (kcal) wurde während des Laufband-Belastungsprotokolls mittels Spirometer softwareseitig („breath-by-breath“) generiert und einer nachgeordneten Analyse in Microsoft Excel (Microsoft Cooperation, 2016) unterzogen.

Zur Messung des Parameters  $\text{VO}_2\text{max}$  ( $\text{ml}/\text{min}/\text{kg}$ ) nutzt die GF5 die Relation zwischen HF und Geschwindigkeit (vgl. Firstbeat Technologies Ltd., 2017), wobei letztgenannte über die Positionsdaten des Nutzers generiert wird (vgl. Garmin Ltd., 2017). Durch die Notwendigkeit der Ermittlung des Positionsparameters war es nicht möglich, die  $\text{VO}_2\text{max}$  im Rahmen des Laufband-Ausbelastungstests (LS 3) anhand der GF5 zu bestimmen, was die nachgeordnete Durchführung des oben beschriebenen Feldtests begründet. Die zur Überprüfung der Reliabilität der GF5 hinsichtlich des Parameters  $\text{VO}_2\text{max}$  in LS 3 über die Referenzmethode SPE

gemessene  $\text{VO}_2\text{max}$  ( $\text{ml}/\text{min}/\text{kg}$ ) wurde unmittelbar mit Beendigung des Laufband-Ausbelastungstests softwareseitig generiert und ausgelesen.

### Missing data

Fehlende Datensätze waren im Verbindungsabbruch zwischen der GF5 und dem Brustgurtsensor in LS 1, der Teilnahmeunfähigkeit eines Probanden aufgrund einer akuten Angina tonsillaris in LS 2 sowie in fehlerhaften, Bluetooth-basierten HF-Übertragungen in LS 3 bzw. fehlgeschlagenen Aufzeichnungen der  $\text{VO}_2\text{max}$  seitens der GF5 im Feldszenario begründet.

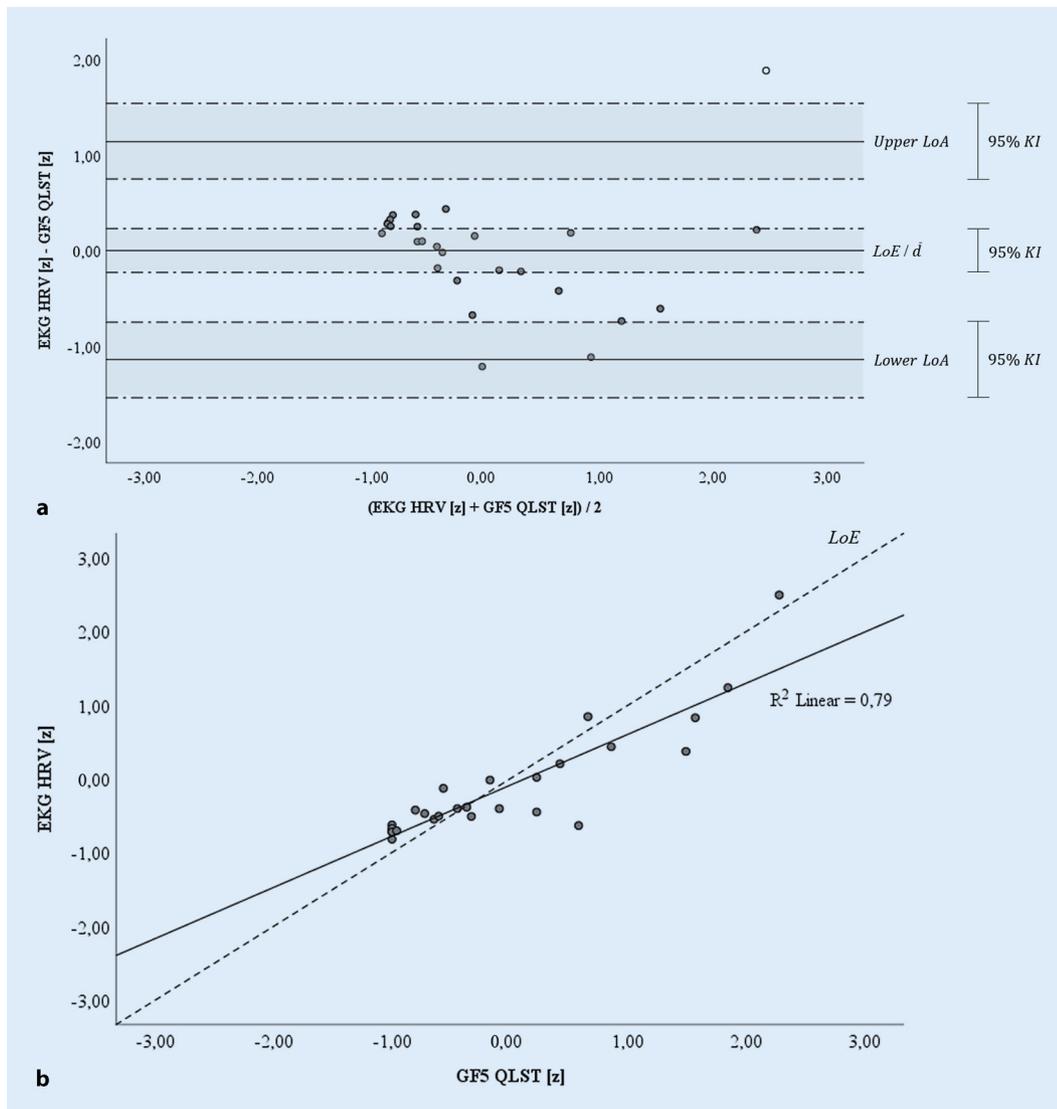
### Datenanalyse

Im ersten Schritt werden die gemittelten Werte aus Referenzmethode und GF5 den Differenzwerten beider Methoden in Form eines Bland-Altman-Plots (B-A-P; Bland & Altman, 1986) gegenübergestellt. Gegen die vermeintlich naheliegende Verwendung der Referenzmethode auf der x-Achse (vgl. Krouwer, 2008) spricht die von Bland und Altman (1995) vorgetragene Argumentation, „A plot of the difference against the standard measurement is sometimes suggested, but this will always appear to show a relation between difference and magnitude when there is none.“ (S. 1085). Im resultierenden B-A-P werden die obere bzw. untere Grenze des 95 % Konfidenzintervalls (KI) der Limits of Agreement (LoA) zur Identifikation von Ausreißerwerten genutzt.

Für den zweiten Schritt, die rechnerische Überprüfung der Reliabilität der GF5 auf Basis der (ausreißerbereinigten) Daten, wurde der Konkordanzkoeffizient nach Lin ( $CCC_{Lin}$ ; Lin, 1989) herangezogen. Im Unterschied zum Produkt-Moment-Korrelationskoeffizienten  $r$  nach Pearson (1896), welcher lediglich eine (populationsbezogene) Bewertung des linearen Assoziationsgrades zwischen beiden Messreihen ermöglicht (vgl. Bland & Altman, 1986), inkludiert der Lin-Korrekturterm über die Präzision (engl.: „precision“)  $p$  hinausgehende Informationen zu additiven (engl.: „location shift“) sowie multiplikativen (engl.: „scale shift“) Skalenunterschie-

den. Durch die Kombination von  $p$  mit den genannten Teilkomponenten der „accuracy“ ( $C_b$ ) entsteht die Möglichkeit zur Quantifizierung der exakten intraindividuellen Konkordanz zwischen zwei Messmethoden. Dieser Vorteil konnte im Falle des standardisierten Stressparameters nicht genutzt werden, da  $C_b$  aufgrund der Standardisierung trivialerweise 1 beträgt; die nachfolgenden Aussagen zur Stressmessung müssen daher auf den Konkordanz-Teilaspekt  $p$  beschränkt bleiben. Die Beurteilung des  $CCC_{Lin}$  erfolgt nach der von Landis und Koch (1977) für Cohens Kappa (Cohen, 1960) vorgeschlagenen Klassifizierung. Neben  $CCC_{Lin}$  wird nachfolgend auch der vermeintlich gebräuchlichere Intraklassen-Korrelationskoeffizient ( $ICC$ , „two-way random effects, absolute agreement, single measurement“;  $ICC [2, 1]$ ) berichtet, welcher numerisch jedoch nahezu identische Ergebnisse liefert. Die Auswahl des  $ICC$  im Kontext der Konkordanzbeurteilung erfolgte in Orientierung an Koo und Li (2016).

Im dritten Schritt werden die B-A-Ps zur Beantwortung der Frage herangezogen, ob und wenn ja, in welchem Maße die GF5 den interessierenden Messparameter im Vergleich zur Referenzmethode (systematisch) über- bzw. unterschätzt. Hierzu dient der Vergleich von mittlerer Differenz ( $\bar{d}$ ) und der, eine komplette Übereinstimmung repräsentierenden, Nulllinie (engl.: „line of equality“, LoE); dieser Vergleich musste aufgrund der Standardisierung der Stresswerte und einer trivialerweise resultierenden mittleren Messwertdifferenz von  $\bar{d} = 0$  auf die Parameter Kalorienverbrauch und  $\text{VO}_2\text{max}$  beschränkt bleiben. Die LoA begrenzen hierbei das Fehlerintervall, in dem, unter der Voraussetzung normalverteilter Differenzen, 95 % aller Messwertdifferenzen ( $\bar{d} \pm 1,96 \text{ SD}$ ) zwischen beiden Methoden inkludiert sind. Die Überprüfung der Normalverteilungsannahme erfolgt sowohl graphisch anhand des Quantil-Quantil(Q-Q)-Diagramms als auch rechnerisch anhand des Shapiro-Wilk-Tests (S-W-T; Shapiro & Wilk, 1965). Die Entscheidung für den S-W-T wird über die Teststärke-Vergleichsstudie von Razali und Wah (2011) begründet. Zur Beantwortung der Frage, ob die



**Abb. 1** ◀ Graphische Darstellung zur partiellen Beurteilung der Konkordanz von Garmin fenix® 5 (GF5) und Elektrokardiographie (EKG) hinsichtlich der Messung des standardisierten Stressparameters (QLST/HRV). **a** Bland-Altman-Plot zur Identifikation von Ausreißerwerten auf Basis der z-standardisierten Messwerte mit mittlerem Methodendifferenzwert ( $\bar{d}$ ), Line of Equality (LoE), Limits of Agreement (LoA;  $Upper\ LoA : \bar{d} + 1,96\ SD$ ;  $Lower\ LoA : \bar{d} - 1,96\ SD$ ) und 95 % Konfidenzintervallen (95 % KI). **b** Streudiagramm zur Beurteilung der Präzision ( $p = 0,89$ ) auf Basis der ausreißerbereinigten, z-standardisierten Messwerte

Messwertdifferenzen systematisch über den gesamten Messbereich der x-Achse streuen, wird das Streuungsmuster der Differenzwerte über den Messbereich im Hinblick auf eine mögliche Verletzung der Homoskedastizitätsannahme beurteilt; zur rechnerischen Verifizierung ebd. Annahme dient der modifizierte Breusch-Pagan-Test (B-P-T; Breusch & Pagan, 1979). Im Falle einer verletzten Normalverteilungsannahme bzw. eines heteroskedastischen Streuungsmusters werden die Original-Messdaten beider Methoden einer logarithmischen Transformation unterzogen; zur besseren Interpretation der resultierenden LoA in der ursprünglichen Messrelation erfolgt eine abschließende Rücktransformation der logarithmierten Messwerte über die

Exponentialfunktion (engl.: „antilog“; vgl. Bland & Altman, 1999) sowie eine darauf gründende, prozentuale Darstellung der ermittelten Abweichungen.

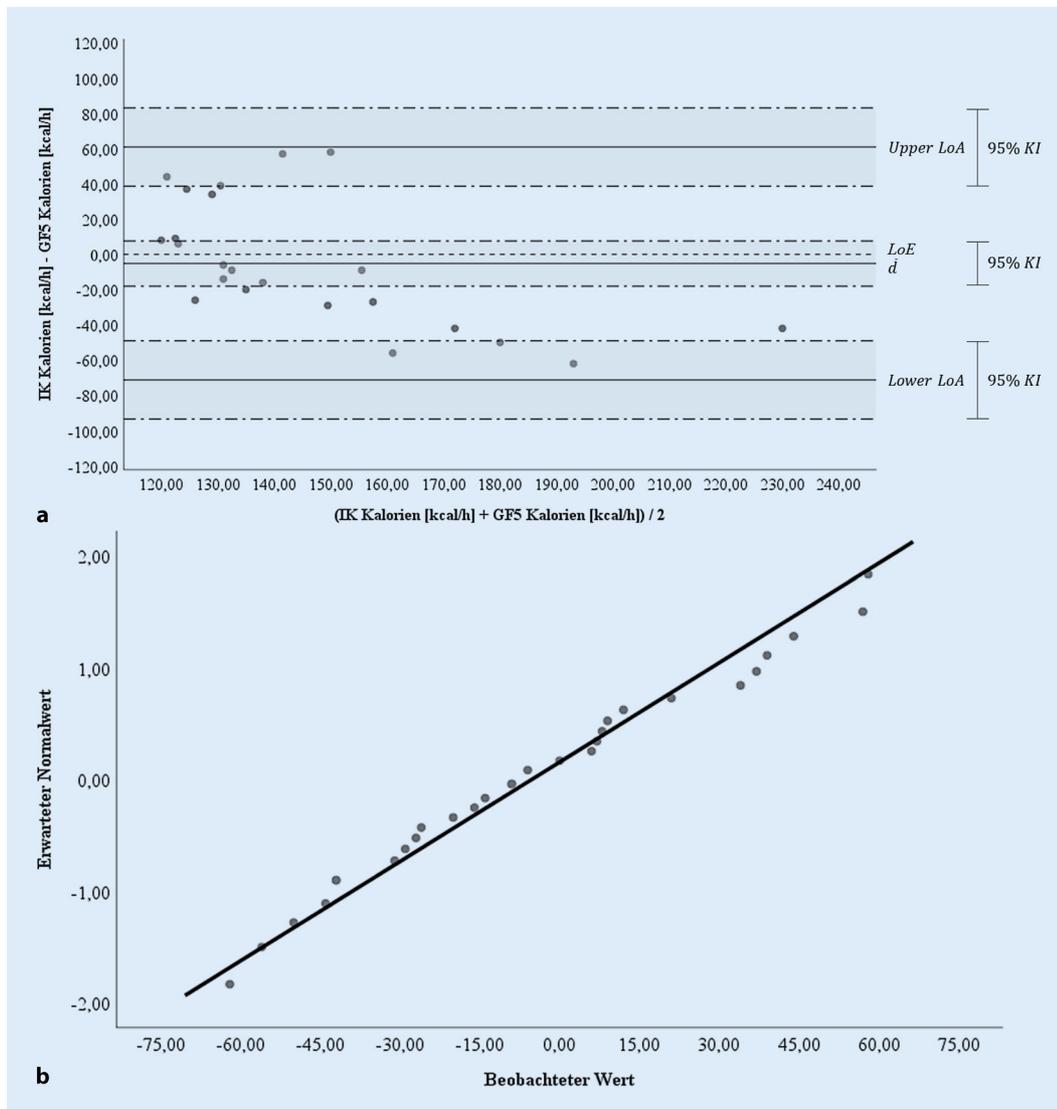
Die hier beschriebenen Analysen wurden mit der Software Statistical Package for the Social Science, Version 25 (IBM Corp., 2017) durchgeführt.

Alle beschriebenen Untersuchungen am Menschen wurden mit Zustimmung der zuständigen Ethik-Kommission, im Einklang mit nationalem Recht sowie gemäß der Deklaration von Helsinki 1975 (in der aktuellen, überarbeiteten Fassung) durchgeführt. Von allen beteiligten Probanden liegt eine Einverständniserklärung vor.

## Ergebnisse

### Stichprobenkennwerte

Nach Anwendung der Ein- und Ausschlusskriterien fanden 30 männliche Probanden mit einem mittleren Alter von 23,13 Jahren ( $SD = 2,50$  Jahre; Range = 9 Jahre; Min. = 20 Jahre; Max. = 29 Jahre) und einem mittleren BMI von  $24,95\ kg/m^2$  ( $SD = 2,45\ kg/m^2$ ; Range =  $8,8\ kg/m^2$ ; Min. =  $21,5\ kg/m^2$ ; Max. =  $30,3\ kg/m^2$ ) Eingang in die Hauptuntersuchung. Aufgrund fehlender Werte konnte zur Überprüfung der Reliabilität der GF5 schlussendlich auf 28 (Stress), 29 (Kalorienverbrauch) bzw. 23 ( $VO_2\ max$ ) vollständige Datensätze zurückgegriffen werden.



**Abb. 2** ◀ Graphische Darstellung zur umfassenden Beurteilung der Konkordanz von Garmin fenix® 5 (GF5) und indirekter Kalorimetrie (IK) hinsichtlich der Messung des Parameters Kalorienverbrauch (Kalorien;  $CCC_{Lin} = 0,43$  [ $p = 0,52$ ,  $C_b = 0,82$ ]). **a** Bland-Altman-Plot zur Beurteilung der Verteilungssystematik der Methodendifferenzwerte (y-Achse) über die Methodenmittelwerte (x-Achse) mit mittlerem Methodendifferenzwert ( $\bar{d}$ ), Line of Equality (LoE), Limits of Agreement (LoA; Upper LoA:  $\bar{d} + 1,96 SD$ ; Lower LoA:  $\bar{d} - 1,96 SD$ ) und 95 % Konfidenzintervallen (95 % KI). **b** Q-Q-Diagramm zur Beurteilung der Normalverteilungsannahme der Methodendifferenzwerte

## Stress

Im Rahmen der Stressexposition in LS 1 erzielten die Probanden ( $N = 28$ ) einen mittleren HF-Wert von  $235,36 \text{ ms}^2$  ( $SD = 269,09 \text{ ms}^2$ ; Range =  $1138,13 \text{ ms}^2$ ; Min. =  $15,07 \text{ ms}^2$ ; Max. =  $1153,20 \text{ ms}^2$ ) sowie einen mittleren QLST-Score von  $74,71$  ( $SD = 25,34$ ; Range =  $83$ ; Min. =  $17$ ; Max. =  $100$ ).

Die Gegenüberstellung der gemittelten Werte aus Referenzmethode und GF5 und den Differenzwerten beider Methoden auf Basis der vorangegangenen z-Standardisierung der Rohwerte zeigt, dass ein Messwert außerhalb des 95 % KI des oberen LoA liegt (vgl. **Abb. 1a**); dieser wurde in der weiteren Analyse der Daten nicht berücksichtigt.

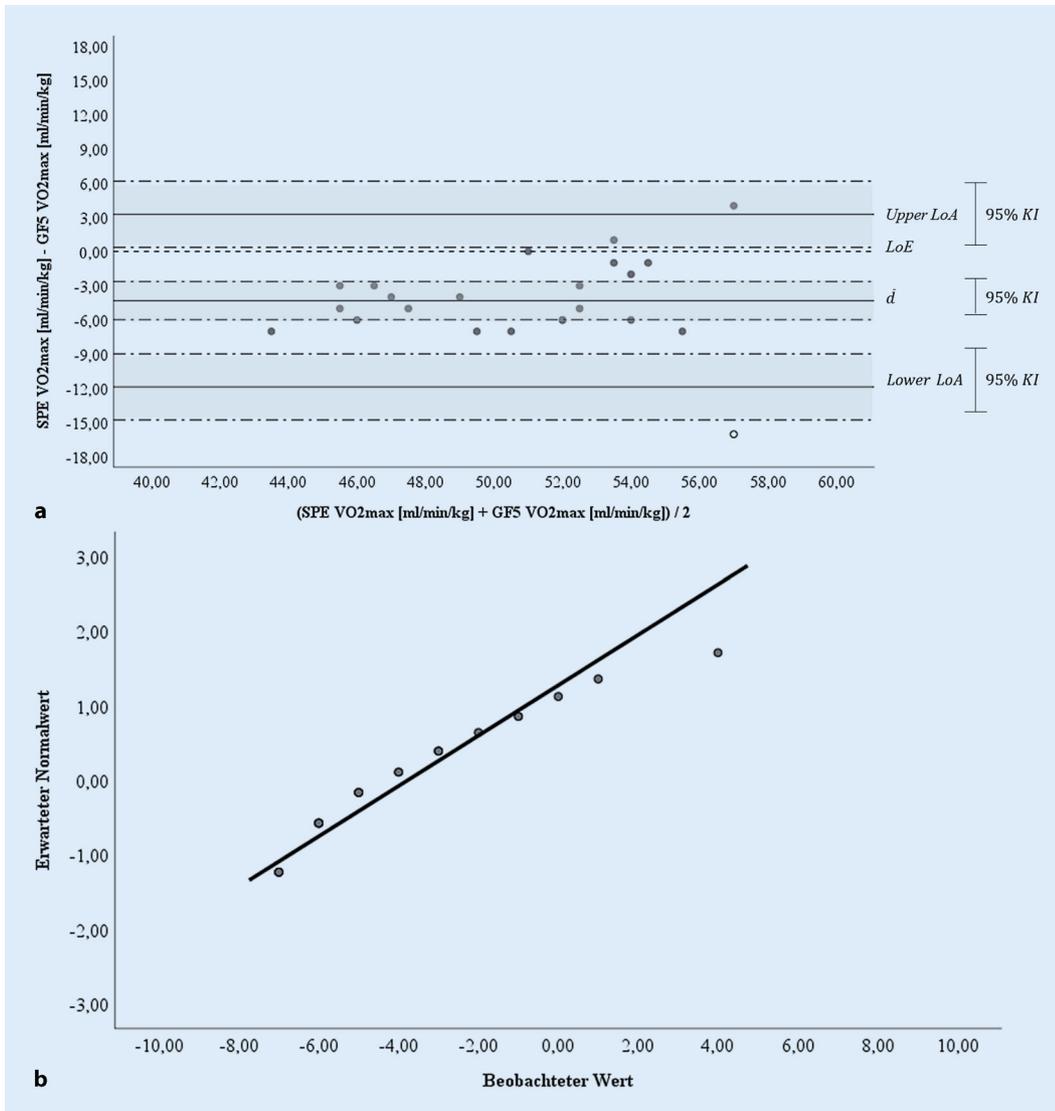
Der rechnerische Vergleich der in LS 1 erhobenen Untersuchungsdaten ( $N = 27$ ) spiegelt eine hohe Präzision der GF5 im Vergleich mit der Referenzmethode EKG hinsichtlich der Messung des z-standardisierten Stressparameters wider ( $p = 0,89$ ; **Abb. 1b**).

## Kalorienverbrauch

Im Rahmen der moderaten Laufbandbelastung in LS 2 erzielten die Probanden ( $N = 29$ ) einen mittleren Kalorienverbrauch von  $140,21 \text{ kcal}$  ( $SD = 21,25 \text{ kcal}$ ; Range =  $99 \text{ kcal}$ ; Min. =  $110 \text{ kcal}$ ; Max. =  $209 \text{ kcal}$ ; IK) bzw. von  $145,41 \text{ kcal}$  ( $SD = 39,52 \text{ kcal}$ ; Range =  $156 \text{ kcal}$ ; Min. =  $99 \text{ kcal}$ ; Max. =  $251 \text{ kcal}$ ; GF5).

Die Gegenüberstellung der gemittelten Werte aus Referenzmethode und GF5 und den Differenzwerten beider Methoden, zeigt, dass alle Messwerte innerhalb der 95 % KI der LoA liegen (**Abb. 2a**). Der rechnerische Vergleich der in LS 2 erhobenen Untersuchungsdaten spiegelt eine mittlere Konkordanz zwischen der GF5 und der Referenzmethode IK hinsichtlich der Messung des Kalorienverbrauchs wider ( $CCC_{Lin} = 0,43$  [ $p = 0,52$ ,  $C_b = 0,82$ ];  $ICC(2, 1) = 0,44$ , 95 % CI [ $0,09$ ;  $0,69$ ]).

Die graphische Analyse des Q-Q-Diagramms spricht für die Annahme näherungsweise normalverteilter Kalorienverbrauch-Messwertdifferenzen (**Abb. 2b**); inferenzstatistische Argumente für die Beibehaltung der, keine



**Abb. 3** ◀ Graphische Darstellung zur umfassenden Beurteilung der Konkordanz von Garmin fenix® 5 (GF5) und Spiroergometrie (SPE) hinsichtlich der Messung des Parameters maximale Sauerstoffaufnahme ( $VO_{2max}$ ;  $CCC_{Lin} = 0,50$  [ $p = 0,77, C_b = 0,66$ ]). **a** Bland-Altman-Plot zur Beurteilung der Verteilungssystematik der Methodendifferenzwerte (y-Achse) über die Methodemittelwerte (x-Achse) mit mittlerem Methodendifferenzwert ( $\bar{d}$ ), Line of Equality (LoE), Limits of Agreement (LoA; Upper LoA:  $\bar{d} + 1,96 SD$ ; Lower LoA:  $\bar{d} - 1,96 SD$ ) und 95 % Konfidenzintervallen (95 % KI). **b** Q-Q-Diagramm zur Beurteilung der Normalverteilungsannahme der Methodendifferenzwerte auf Basis der ausreißerbereinigten Messwerte

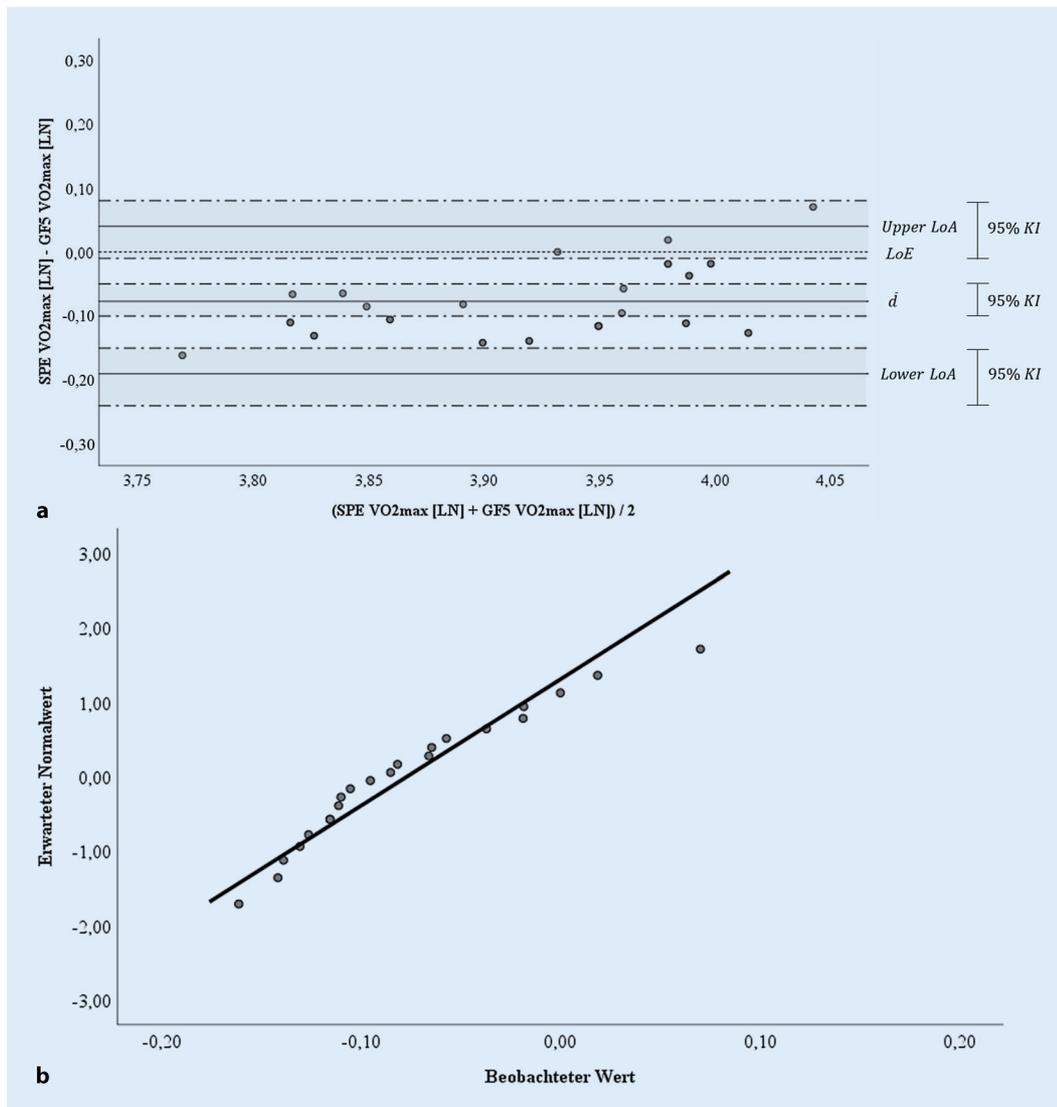
überzufällige Verletzung der Normalverteilung implizierenden, Nullhypothese finden sich im S-W-T [ $W(29) = 0,97$ ;  $p = 0,520$ ]. Die Verteilung der Differenzwerte im B-A-P (Abb. 2a) befördert die Homoskedastizitätsannahme; für die Beibehaltung der, keine systematische Verteilung der Differenzwerte über den Messbereich implizierenden, Nullhypothese spricht das Ergebnis des modifizierten B-P-T [ $\chi^2(1) = 0,15$ ;  $p = 0,695$ ]. Auf Grundlage der normalverteilten und homoskedastischen Differenzwerte wird eine mittlere Messwertdifferenz (SPE-GF5) von  $\bar{d} = -5,21$  kcal ( $SD = 33,71$  kcal; 95 % KI [-18,03 kcal; +7,62 kcal]) bestimmt, wobei die LoE im 95 % KI ebd. liegt und sich 95 % der Messwertdifferenzen in einem LoA-Intervall zwischen -71,29 kcal

(95 % KI [-93,46 kcal; -49,11 kcal]; Lower LoA) und +60,87 kcal (95 % KI [+38,70 kcal; +83,04 kcal]; Upper LoA) befinden (Abb. 2a); dies bedeutet, dass die GF5 den Kalorienverbrauch im Mittel zufällig um -1,48 % (95%KI [-10,03 %; +7,07 %]) überschätzt und das mittlere Abweichungsspektrum der GF5 von -45,52 % (95 % KI [-60,30 %; -30,74 %]) Überschätzung bis zu +42,56 % (95 % KI [+27,79 %; +57,34 %]) Unterschätzung des wahren Kalorienverbrauchs reicht. Unter Einbeziehung der LoA-Konfidenzintervalle ist dabei herauszustellen, dass die maximalen Abweichungen im ungünstigsten Replikationsfall zwischen -93,46 kcal (-60,30 %) Überschätzung und +83,04 kcal (+57,34 %) Unterschätzung liegen.

### VO<sub>2max</sub>

Im Rahmen der Laufausbelastung in LS 3 bzw. dem Feldszenario erzielten die Probanden ( $N = 23$ ) eine mittlere  $VO_{2max}$  von 48,70 ml/min/kg ( $SD = 4,50$  ml/min/kg; Range = 19 ml/min/kg; Min. = 40 ml/min/kg; Max. = 59 ml/min/kg; Spiroergometrie) bzw. 53,00 ml/min/kg ( $SD = 4,23$  ml/min/kg; Range = 18 ml/min/kg; Min. = 47 ml/min/kg; Max. = 65 ml/min/kg; GF5).

Die Gegenüberstellung der gemittelten Werte aus Referenzmethode und GF5 und den Differenzwerten beider Methoden, zeigt, dass ein Messwert außerhalb des 95 % KI des unteren LoA liegt (Abb. 3a); dieser wurde in der weiteren Analyse der Daten nicht be-



**Abb. 4** ◀ Graphische Darstellung zur umfassenden Beurteilung der Konkordanz von Garmin fenix® 5 (GF5) und Spiroergometrie (SPE) hinsichtlich der Messung des Parameters *maximale Sauerstoffaufnahme* ( $VO_{2max}$ ;  $CCC_{Lin} = 0,50$  [ $p = 0,77$ ,  $C_b = 0,66$ ]). **a** Bland-Altman-Plot zur Beurteilung der Verteilungssystematik der Methodendifferenzwerte (y-Achse) über die Methodenmittelwerte (x-Achse) auf Basis der logarithmierten Messwerte mit mittlerem Methodendifferenzwert ( $\bar{d}$ ), Line of Equality (LoE), Limits of Agreement (LoA; *Upper LoA*:  $\bar{d} + 1,96 SD$ ; *Lower LoA*:  $\bar{d} - 1,96 SD$ ) und 95 % Konfidenzintervallen (95 % KI). **b** Q-Q-Diagramm zur Beurteilung der Normalverteilungsannahme auf Basis der logarithmierten Messwerte

rücksichtigt. Der rechnerische Vergleich der in LS 2 und dem Feldszenario erhobenen Untersuchungsdaten ( $N = 22$ ) spiegelt eine mittlere Konkordanz zwischen der GF5 und der Referenzmethode SPE hinsichtlich der Messung der maximalen Sauerstoffaufnahme wider ( $CCC_{Lin} = 0,50$  [ $p = 0,77$ ,  $C_b = 0,66$ ];  $ICC(2, 1) = 0,52$ , 95 % CI  $[-0,09; +0,82]$ ).

Die graphische Analyse des Q-Q-Diagramms spricht für die Annahme nichtnormalverteilter  $VO_{2max}$ -Messwertdifferenzen (Abb. 3b); dahingehende inferenzstatistische Argumente für die Annahme der, eine überzufällige Verletzung der Normalverteilung implizierenden, Alternativhypothese finden sich im S-W-T [ $W(22) = 0,90$ ;  $p = 0,035$ ]. Die (ausreißerbereinigte) Verteilung der

Differenzwerte im B-A-P (Abb. 3a) befördert die Heteroskedastizitätsannahme; für die Annahme der, eine systematische Verteilung der Differenzwerte über den Messbereich implizierenden, Alternativhypothese spricht das Ergebnis des modifizierten B-P-T [ $\chi^2(1) = 7,62$ ;  $p = 0,006$ ].

Die logarithmische Transformation beider Messreihen bewirkt eine günstigere Verteilung der Messwertdifferenzen dahingehend, dass die, keine überzufällige Verletzung der Normalverteilungsannahme implizierende, Nullhypothese hier beibehalten werden kann [ $W(22) = 0,94$ ;  $p = 0,185$ ; Abb. 4b]. Stochastische Argumente für die gleichsame Beibehaltung der, keine systematische Verteilung der logarithmierten Diffe-

renzwerte über den Messbereich implizierenden, Nullhypothese, finden sich dagegen trotz leicht verbesserter Streuungseigenschaften nicht [ $\chi^2(1) = 6,20$ ;  $p = 0,013$ ; Abb. 4a].

Auf Grundlage der *normalverteilten* – weiterhin jedoch leicht heteroskedastischen – logarithmierten Differenzwerte wird eine mittlere Messwertdifferenz (SPE-GF5) von  $\bar{d} = -0,08$  ( $SD = 0,06$ ; 95 % KI  $[-0,10; -0,05]$ ) Einheiten bestimmt, wobei die LoE außerhalb des 95 % KI ebd. liegt und sich 95 % der Messwertdifferenzen in einem LoA-Intervall zwischen  $-0,19$  (95 % KI  $[-0,24; -0,15]$ ; *Lower LoA*) und  $+0,04$  (95 % KI  $[-0,01; +0,08]$ ; *Upper LoA*) Einheiten befinden (Abb. 4a). Die Rücktransformation der logarithmierten Daten über

die Umkehrfunktion führt zu einer mittleren Messwertdifferenz von  $\bar{d} = +0,93$  (95 % KI [+0,90; +0,95]) Einheiten in einem LoA-Intervall zwischen +0,83 (95 % KI [+0,79; +0,86]; *Lower LoA*) und +1,04 (95 % KI [+0,99; +1,09]; *Upper LoA*) Einheiten; dies bedeutet, dass die GF5 die Referenzmethode überzufällig um -7,40 % (95 % KI [-9,78 %; -4,96 %]) überschätzt und das mittlere Abweichungsspektrum der GF5 von -17,49 % (95 % KI [-21,12 %; -13,68 %]) Überschätzung bis zu +3,92 % (95 % KI [-0,66 %; +8,71 %]) Unterschätzung der wahren  $VO_{2max}$  reicht. Unter Einbeziehung der LoA-Konfidenzintervalle ist dabei herauszustellen, dass die maximalen Abweichungen im ungünstigsten Replikationsfall zwischen +0,79 Einheiten (-21,12 %) Überschätzung und +1,09 Einheiten (+8,71 %) Unterschätzung liegen.

## Diskussion

### Zusammenfassung, Einordnung und praktische Implikation

Gegenstand der vorliegenden Untersuchung war die Überprüfung der Reliabilität der als marktrelevant identifizierten Premium Multisport-Smartwatch GF5 anhand des Teilaspektes der Konkordanz und hinsichtlich der Parameter Stress, Kalorienverbrauch und  $VO_{2max}$ . Eine umfassende Überprüfung der exakten intraindividuellen Konkordanz der GF5 im Vergleich mit der Referenzmethode EKG hinsichtlich der Messung des Stressparameters war aufgrund unterschiedlicher Messmetriken nicht möglich; die realisierbaren rechnerischen Analysen im Bereich der Stressmessung sprechen jedoch zumindest für eine hohe Präzision  $p$  der GF5. Der in der Zusammenfassung des systematischen Reviews von Georgiou et al. (2018) formulierten Einschätzung „Wearable devices may provide a promising alternative solution for measuring [H]RV.“ (S. 7) kann auf Basis der vorliegenden Befunde somit nicht grundsätzlich widersprochen werden. Ferner zeigen die rechnerischen Analysen eine forschungshypothesenkonform mittlere Konkordanz der GF5 im Vergleich mit der jeweiligen Referenzmethode

IK bzw. SPE hinsichtlich der Messung der Parameter Kalorienverbrauch bzw.  $VO_{2max}$ . Das Ausmaß der Methoden(in)konkordanz kann dabei über die graphisch abgeleitete, mittlere Messwertdifferenz beurteilt werden. Hier zeigen die vorliegenden Befunde im Mittel eine gerade 1,48-prozentige, zufällige Überschätzung des Kalorienverbrauchs der GF5 im Vergleich zur Referenzmethode; die vorliegenden Befunde spiegeln im Betrag eine – wenngleich nicht generalisierbare – geringere mittlere prozentuale Abweichung als die im Firstbeat-Whitepaper über den Mean Absolute Percentage Error (MAPE) zusammengefassten Befunde (MAPE = 7–11 %; vgl. Firstbeat Technologies Ltd., 2012) und sprechen somit zunächst für die Reliabilität der GF5. Diese Einschätzung ist zu relativieren, zieht man zur Beurteilung die im Rahmen der graphischen Analyse bestimmten LoA heran. Hiernach muss davon ausgegangen werden, dass das mittlere Abweichungsspektrum der GF5 von -45,52 % Überschätzung bis zu +42,56 % Unterschätzung des wahren Kalorienverbrauchs reicht. Praktisch und unter Einbeziehung der Metabolic Equivalent of Task (MET)-Aktivitätsentsprechungen von Ainsworth et al. (2011) bedeutet dies, dass die GF5 für eine Person mit 80 kg Körpergewicht bei einem Dauerlauf mit 6,4 km/h (6 METs) anstelle eines zu erwartenden Kalorienverbrauchs von 480 kcal/h, einen Kalorienverbrauch von maximal 698,50 kcal/h ( $\rightarrow$  8,73 METs) oder minimal 275,71 kcal/h ( $\rightarrow$  3,44 METs) ausgeben könnte; nach diesen Berechnungen würde die moderate Laufbelastung von Seiten der GF5 energetisch somit zumindest potenziell entweder einer schnelleren Laufeinheit mit 8,36 km/h (9,0 METs) oder aber einem Spaziergang mit leichtem Tragegewicht (3,0 METs) gleichgesetzt. Ein ähnliches Bild ergibt sich bei der Einordnung der Messwertdifferenzen im Bereich der  $VO_{2max}$ . Hier zeigen die vorliegenden Befunde, dass die GF5 die  $VO_{2max}$  im Mittel überzufällig um -7,40 % überschätzt und damit im Betrag nur 2,4 % über dem im Firstbeat-Whitepaper (vgl. Firstbeat Technologies Ltd., 2017) berichteten MAPE liegt; auch dieser Befund kann zunächst als Argument

für die Reliabilität der GF5 aufgefasst werden. Diese Einschätzung ist jedoch erneut zu relativieren, zieht man zur Beurteilung die im Rahmen der graphischen Analyse bestimmten LoA heran. Hiernach muss davon ausgegangen werden, dass das mittlere Abweichungsspektrum der GF5 von -17,49 % Überschätzung bis zu +3,92 % Unterschätzung der wahren  $VO_{2max}$  reicht. Nach den seitens der GF5 referenzierten  $VO_{2max}$ -Standard-Bewertung des Cooper-Institutes (2007) bedeutet dies, dass die GF5 für einen männlichen Probanden zwischen 20 und 29 Jahren mit einer wahren  $VO_{2max}$  von 47,25 ml/min/kg, eine  $VO_{2max}$  von maximal 55,51 ml/min/kg oder minimal 45,40 ml/min/kg ausgeben könnte; seine eigentlich als „good“ zu klassifizierende  $VO_{2max}$  würde von Seiten der GF5 somit zumindest potenziell entweder als „excellent to superior“ (Überschätzung) oder „fair to good“ (Unterschätzung) klassifiziert. In Anbetracht dieser Klassifikationsstreuung kann die gleichsam querschnittlich basierte und herstellernabhängige Einschätzung von Anderson et al. (2019), „[...] the GF5 can be an option for a person seeking an affordable and easily available method of determining  $VO_{2max}$ . (S. 147)“ nicht vorbehaltlos geteilt werden.

Dass die aufgezeigte Konkordanzsituation der GF5 durchaus auch auf andere marktrelevante WD übertragbar ist, zeigen die aktuellen Daten von Passler et al. (2019). Die in der vergleichenden Arbeit von Düking et al. (2020) ermittelten und zum Teil durchaus ermutigenden linearen Befunde zur Validität ebensolcher WD sollten im Hinblick auf den hier explizierten Reliabilitätsaspekt der Konkordanz (aufgrund der abweichenden methodisch-analytischen Zuganges) in jedem Falle nicht hypothesenleitend verwendet werden.

## Limitationen

Bei der vorgängigen Interpretation der Befunde sind sowohl auf die Stichprobe als auch auf den Untersuchungsablauf bezogene sowie analytische und interpretative Limitationen zu berücksichtigen.

Der analysierte Stichprobenumfang lag in der vorliegenden Untersuchung

gerade zwischen 22 (VO<sub>2</sub>max) und 29 (Kalorienverbrauch) Probanden. Die aus der geringen Fallzahl resultierenden, weiten Konfidenzintervalle bewirken, dass das Ergebnis einer im Mittel geringen Überschätzung der GF5 im Parameter Kalorienverbrauch interpretativ auf die untersuchte Stichprobe begrenzt bleiben sollte. Ferner führen die weiten Konfidenzintervalle bei der im Ergebnisteil abschließend vorgenommenen Interpretation auf Basis der maximal möglichen Abweichungen zu einer potenziell negativen Überzeichnung der tatsächlichen Messgenauigkeit der GF5 hinsichtlich der Parameter Kalorienverbrauch und VO<sub>2</sub>max an den unteren bzw. oberen Rändern.

Weitere Einschränkungen der Befundinterpretation liegen im Untersuchungsablauf begründet. Hier ist zu berücksichtigen, dass die Probanden das bereitgestellte GF5-Modell erstmalig und einzig innerhalb der gegebenen Szenarien genutzt haben. Hierdurch wurde die in der GF5 inkludierte Möglichkeit der Verbesserung der Messgenauigkeit des Stresswertes (vgl. Firstbeat Technologies Ltd., 2020), des Kalorienverbrauchs über ein neuronales Netz (vgl. Firstbeat Technologies Ltd., 2012) sowie der maximalen Sauerstoffaufnahme (vgl. Garmin Ltd, 2017) ausgeschlossen und die hier interessierende Methodenkonkordanz im Vergleich zu einem stärker nutzeradaptierten GF5-Modell somit potenziell unterschätzt. Ferner ist zu berücksichtigen, dass die VO<sub>2</sub>max in der vorliegenden Untersuchung an zwei unterschiedlichen Testtagen erfasst wurde und das aufgrund der Herstellerempfehlung umgesetzte Constant-load-Protokoll nicht dem etablierten und gleichsam umgesetzten Rampenprotokoll entsprach. Unter der Prämisse einer VO<sub>2</sub>max-Variation in Abhängigkeit u. a. von Ermüdungs- und Ernährungszustand bzw. Protokolllänge kann die gerade mittlere VO<sub>2</sub>max-Konkordanz der GF5 im Vergleich zur Referenzmethode SPE somit nicht ausschließlich Device-seitig begründet werden.

In Bezug auf die Interpretation der Methodenkonkordanz im Bereich Stress ist aus analytischer Perspektive neuerlich zu betonen, dass die für den Stresspa-

rameter notwendigerweise erfolgte Beschränkung auf  $p$  keinen Schluss auf die exakte intraindividuelle Konkordanz der GF5 ermöglicht. Ferner ist in Bezug auf die Interpretation der Methodenkonkordanz im Bereich der VO<sub>2</sub>max zu berücksichtigen, dass diese durch einen vergleichsweise extremen Wert im oberen Messspektrum mitbestimmt wurde und die Messabweichungen in den unteren VO<sub>2</sub>max-Bereichen in engeren LoA vermutet werden können. Auf eine Entgegnung des leicht heteroskedastischen Streuungsmusters mit einer bereichsspezifischen Grenzwertanalyse (vgl. Watson & Petrie, 2010) wurde aufgrund der insgesamt geringen Fallzahlen hier jedoch verzichtet.

Bei der Interpretation der im Ergebnisteil vorangestellten, rechnerischen Befunde zur Methodenkonkordanz ist abschließend zu berücksichtigen, dass die getroffenen Einschätzungen zur Einordnung des CCC<sub>Lim</sub> auf einem zwar etablierten, jedoch eher moderaten Klassifikationsvorschlag für strenggenommen kategoriale Variablen gründen.

## Forschungsperspektiven

Der bei der Analyse der Methodenkonkordanz im Parameter Kalorienverbrauch aufgezeigten Problematik der zu geringen Teststärke im Hinblick auf eine populationsbezogene Beurteilung der mittleren Überschätzung sollte in der notwendigen Replikation der vorliegenden Primärstudie mit einem optimalen Stichprobenumfang begegnet werden; durch entsprechend verkleinerte Konfidenzintervalle kann dann auch von einer Präzisierung der Befundlage an den Rändern ausgegangen werden. Dahingehende kalkulatorische Empfehlungen im Kontext von „reproducibility studies“ finden sich bei Watson und Petrie (2010).

Bei der Konzeption zukünftiger Untersuchungen ist der Beantwortung der Frage, ob und wenn ja, in welchem Maße die Konkordanz zwischen GF5 und Referenzmethode durch eine kontinuierliche Nutzung der GF5 verbessert werden kann, besondere Bedeutung beizumessen. Im Falle einer akzeptablen Annäherung der Lernkurve der GF5 an den wahren Wert stellt sich dann schlussend-

lich die Frage nach der Wiederholbarkeit (engl.: „repeatability“) der Befunde; auch hier finden sich entsprechende kalkulatorische Empfehlungen bei Watson und Petrie (2010). Darüber hinaus ist anzudenken, den Kalorienverbrauch bei unterschiedlichen Laufgeschwindigkeiten und die VO<sub>2</sub>max im Rahmen einer einzigen Testung und somit auf der Basis nur eines Protokolls zu ermitteln.

Für die Datenanalyse im Kontext der notwendigen Replikationsstudien wird der CCC<sub>Lim</sub> aufgrund der Möglichkeit zur Bestimmung der exakten intraindividuellen Konkordanz empfohlen, wobei die gefundenen Effekte anhand eines angemessenen und ggf. restriktiveren (z. B. McBride, 2005) Klassifikationssystems beurteilt werden sollten. Die Frage nach der exakten intraindividuellen Konkordanz zwischen GF5 und Referenzmethode EKG hinsichtlich der Erfassung des Parameters Stress lässt sich dabei erst dann adäquat – d. h., unter Einbeziehung von  $p$  und  $C_b$  – beantworten, wenn der Hersteller den Zugang auf die zu Grunde liegenden HRV-Daten offenlegt.

## Fazit

Die hier ermittelte Präzision der GF5 hinsichtlich der Messung des Stressparameters ist positiv zu sehen; in Ermangelung der Möglichkeit einer umfassenden Konkordanzüberprüfung kann die GF5 aus ebd. Perspektive aktuell jedoch nicht vorbehaltlos zur Stressmessung empfohlen werden. Gleichsam nicht vorbehaltlos, argumentativ jedoch umfassender begründbar, fällt die Empfehlung zum Einsatz der GF5 in den Bereichen Aktivität und Fitness aus. So ist in Anbetracht der hier berichteten mittleren Konkordanz der GF5 in den Parametern Kalorienverbrauch und VO<sub>2</sub>max eine basale Standortbestimmung sowie eine darauf aufbauende gesundheitsorientierte Trainingssteuerung mit dem Ziel einer Gewichtsreduktion und/oder Leistungsentwicklung grundsätzlich denkbar; (gesundheitsorientierte) Nutzer dieses Devices sollten sich dabei jedoch bewusst sein, dass sich gerade bei erstmaliger Verwendung und in Einzelfällen durchaus bedeutsame Abweichungen im Vergleich zum wahren Wert in den Para-

metern Kalorienverbrauch und  $VO_{2max}$  ergeben können. Auch und gerade für ambitionierte (Ausdauer-)Athleten, die den Aspekten der Trainingssteuerung und Leistungskontrolle eine vergleichsweise höhere Gewichtung beimessen (sollten), stellt die GF5 bei erstmaliger Verwendung demnach noch kein hinreichend konkordantes Messinstrument im Vergleich zu einer laborbasierten trainingswissenschaftlichen Diagnostik.

Mit einer verlässlichen Messung steigt die Chance auf ein entsprechendes Kompetenzerleben im Trainingsprozess; dies kann die Entwicklung einer positiven sport- und bewegungsbezogenen Selbstwirksamkeit befördern und somit über den motivationspsychologischen Prozess der Internalisierung der Gefahr eines frühen Drop-outs entgegenwirken (vgl. Je-kauc et al., 2015).

Für die testtheoretische Perspektive des Themenfeldes Bewegung, Gesundheit und Digitalisierung bedeutet dies, dass ein weiteres Investment in die Verbesserung der Reliabilität (und Validität) von WD zu einer dauerhaften Bindung an ressourcenstärkende und anforderungsmindernde körperliche Aktivitäten beitragen kann (vgl. Mercer, Li, Giangregorio, Burns, & Grindrod, 2016), wodurch der einzelne Nutzer besser auf die Herausforderungen des digitalen gesellschaftlichen Wandels vorbereitet wird.

In Anbetracht der Geschwindigkeit der technischen Progression sowie der daraus resultierenden Marktbreite und -heterogenität (Bunn, Navalta, Fontaine, & Reece, 2018; Duffy & Colon, 2019) liegt es in der Verantwortung der Forschenden im Themenfeld Bewegung, Gesundheit und Digitalisierung, diese Entwicklungsschritte durch regelmäßige Untersuchungen wie die vorliegende sowie daraus erarbeitete, systematische Übersichten (u. a. Düking, Hotho, Holmberg, Fuss, & Sperlich, 2016) kritisch zu reflektieren.

## Korrespondenzadresse



**Univ.-Prof. Dr. phil. habil. Matthias Wagner**

Fakultät für Humanwissenschaften, Institut für Sportwissenschaft, Forschungs- und Lehrgebiet Sportpsychologie, Universität der Bundeswehr München Werner-Heisenbergweg 39, 85577 Neubiberg, Deutschland  
matthias.wagner@unibw.de

**Funding.** Open Access funding enabled and organized by Projekt DEAL.

## Einhaltung ethischer Richtlinien

**Interessenkonflikt.** M. Wagner, F. Engel, K. Klier, S. Klughardt, F. Wallner und A. Wiecek geben an, dass kein Interessenkonflikt besteht.

Für diesen Beitrag wurden von den Autoren keine Studien an Menschen oder Tieren durchgeführt. Für die aufgeführten Studien gelten die jeweils dort angegebenen ethischen Richtlinien.

**Open Access.** Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

## Literatur

- Ainsworth, B. E., Haskell, W. L., Herrmann, S. D., Meckes, N., Bassett, D. R., Tudor-Locke, C., Leon, A. S., et al. (2011). 2011 compendium of physical activities: a second update of codes and MET values. *Medicine and Science in Sports and Exercise*, 43(8), 1575–1581. <https://doi.org/10.1249/MSS.0b013e31821ece12>.
- Anderson, J. C., Chisenall, T., Tolbert, B., Ruffner, J., Whitehead, P. N., & Conners, R. T. (2019). Validating the Commercially Available Garmin

- Fenix 5x Wrist-Worn Optical Sensor for Aerobic Capacity. *International Journal for Innovation Education and Research*, 7(1), 147–158. <https://doi.org/10.31686/ijer.Vol7.Iss1.1293>.
- Arojanam, G., Manivannan, N., & Harrison, D. (2019). Review on wearable technology sensors used in consumer sport applications. *Sensors*. <https://doi.org/10.3390/s19091983>.
- Attig, C., & Franke, T. (2019). I track, therefore I walk – Exploring the motivational costs of wearing activity trackers in actual users. *International Journal of Human-Computer Studies*, 127, 211–224. <https://doi.org/10.1016/j.ijhcs.2018.04.007>.
- Banegas, J. R., Ruilope, L. M., La Sierra, A., de Vinyoles, E., Gorostidi, M., de La Cruz, J. J., Williams, B., et al. (2018). Relationship between clinic and ambulatory blood-pressure measurements and mortality. *The New England Journal of Medicine*, 378(16), 1509–1520. <https://doi.org/10.1056/NEJMoa1712231>.
- Barchard, K. A. (2012). Examining the reliability of interval level data using root mean square differences and concordance correlation coefficients. *Psychological Methods*, 17(2), 294–308. <https://doi.org/10.1037/a0023351>
- Becker, P. (1992). Seelische Gesundheit als protektive Persönlichkeitseigenschaft. *Zeitschrift Für Klinische Psychologie*, 21, 64–75.
- Becker, P., Bös, K., & Woll, A. (1994). Ein Anforderungs-Ressourcen-Modell der körperlichen Gesundheit: Pfadanalytische Ueberprüfungen mit latenten Variablen. *Zeitschrift Für Gesundheitspsychologie*, 2(1), 25–48.
- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1(8476), 307–310.
- Bland, J. M., & Altman, D. G. (1995). Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet*, 346(8982), 1085–1087. [https://doi.org/10.1016/S0140-6736\(95\)91748-9](https://doi.org/10.1016/S0140-6736(95)91748-9).
- Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2), 135–160. <https://doi.org/10.1177/096228029900800204>.
- Borg, G. (1998). *Borg's perceived exertion and pain scales*. Champaign: Human Kinetics.
- Brehm, W., Bös, K., Graf, C. H., Hartmann, H., Pahmeier, I., Pfeifer, K., Wagner, P., et al. (2013). Sport als Mittel in Prävention, Rehabilitation und Gesundheitsförderung. Eine Expertise [Sport as a means to prevention, rehabilitation, and health promotion. An expert opinion. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, 56(10), 1385–1389.
- Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5), 1287–1294. <https://doi.org/10.2307/1911963>.
- Bunn, J. A., Navalta, J. W., Fontaine, C. J., & Reece, J. D. (2018). Current state of commercial wearable technology in physical activity monitoring 2015–2017. *International Journal of Exercise Science*, 11(7), 503–515.
- Busch, L., Utesch, T., Bürkner, P.-C., & Strauss, B. (2020). The influence of fitness-app usage on psychological well-being and body awareness—a daily diary randomized trial. *Journal of Sport & Exercise Psychology*. <https://doi.org/10.1123/jsep.2019-0315>.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-mul-

- timethod matrix. *Psychological Bulletin*, 56(2), 81–105. <https://doi.org/10.1037/h0046016>.
- Chung, A. E., Skinner, A. C., Hasty, S. E., & Perrin, E. M. (2017). Tweeting to health: a novel mhealth intervention using Fitbits and twitter to foster healthy lifestyles. *Clinical Pediatrics*, 56(1), 26–32. <https://doi.org/10.1177/0009922816653385>.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior. Perspectives in social psychology*. New York: Plenum.
- Dishman, R. K., & Buckworth, J. (1996). Increasing physical activity: a quantitative synthesis. *Medicine and Science in Sports and Exercise*, 28(6), 706–719. <https://doi.org/10.1097/00005768-199606000-00010>.
- Duffy, J., & Colon, A. (2019). *The best fitness trackers of 2020: PC Reviews*. Verfügbar unter <https://uk.pcmag.com/fitness-trackers/159/the-best-fitness-trackers>. Zugegriffen: 19. Okt. 2020.
- Düking, P., Fuss, F. K., Holmberg, H.-C., & Sperlich, B. (2018). Recommendations for assessment of the reliability, sensitivity, and validity of data provided by wearable sensors designed for monitoring physical activity. *JMIR MHealth and UHealth*, 6(4), e102. <https://doi.org/10.2196/mhealth.9341>.
- Düking, P., Giessing, L., Frenkel, M. O., Koehler, K., Holmberg, H.-C., & Sperlich, B. (2020). Wrist-worn wearables for monitoring heart rate and energy expenditure while sitting or performing light-to-vigorous physical activity: validation study. *JMIR MHealth and UHealth*, 8(5), e16716.
- Düking, P., Hotho, A., Holmberg, H.-C., Fuss, F. K., & Sperlich, B. (2016). Comparison of non-invasive individual monitoring of the training and health of athletes with commercially available wearable technologies. *Frontiers in Physiology*, 7, 71. <https://doi.org/10.3389/fphys.2016.00071>.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132. <https://doi.org/10.1146/annurev.psych.53.100901.135153>.
- Firstbeat Technologies Ltd. (2012). *An energy expenditure estimation method based on heart rate measurement*. [https://assets.firstbeat.com/firstbeat/uploads/2015/10/white\\_paper\\_energy\\_expenditure\\_estimation.pdf](https://assets.firstbeat.com/firstbeat/uploads/2015/10/white_paper_energy_expenditure_estimation.pdf). Zugegriffen: 19. Okt. 2020.
- Firstbeat Technologies Ltd. (2014). *Stress and recovery analysis method based on 24-hour heart rate variability*. [https://assets.firstbeat.com/firstbeat/uploads/2015/10/Stress-and-recovery\\_white\\_paper\\_20145.pdf](https://assets.firstbeat.com/firstbeat/uploads/2015/10/Stress-and-recovery_white_paper_20145.pdf). Zugegriffen: 19. Okt. 2020.
- Firstbeat Technologies Ltd. (2017). *Automated fitness level (VO2max) estimation with heart rate and speed data*. [https://assets.firstbeat.com/firstbeat/uploads/2017/06/white\\_paper\\_VO2max\\_30.6.2017.pdf](https://assets.firstbeat.com/firstbeat/uploads/2017/06/white_paper_VO2max_30.6.2017.pdf). Zugegriffen: 19. Okt. 2020.
- Firstbeat Technologies Ltd. (2020). *Quick stress level test*. <https://www.firstbeat.com/en/consumer-feature/quick-stress-level-test/>. Zugegriffen: 19. Okt. 2020.
- Garmin Ltd (2017). *Garmin fenix®5-Benutzerhandbuch*. [http://static.garmin.com/pumac/fenix5S\\_OM\\_DE.pdf](http://static.garmin.com/pumac/fenix5S_OM_DE.pdf). Zugegriffen: 21. Okt. 2020.
- Georgiou, K., Larentzakis, A. V., Khamis, N. N., Alshuhai-bani, G. I., Alaska, Y. A., & Giallafa, E. J. (2018). Can wearable devices accurately measure heart rate variability? A systematic review. *Folia Medica*, 60(1), 7–20. <https://doi.org/10.2478/folmed-2018-0012>.
- Gerber, M. (2014). *Pädagogische Psychologie im Sportunterricht: Ein Lehrbuch in 14 Lektionen*. Sportwissenschaft studieren, Bd. 9. Aachen: Meyer & Meyer.
- Hills, A. P., Mokhtar, N., & Byrne, N. M. (2014). Assessment of physical activity and energy expenditure: an overview of objective measures. *Frontiers in Nutrition*. <https://doi.org/10.3389/fnut.2014.00005>.
- Hohmann, A., Lames, M., & Letzelter, M. (2014). *Einführung in die Trainingswissenschaft (6., unveränderte Auflage)*. Wiebelsheim: Limpert Verlag.
- Hollmann, W., & Strüder, H. K. (2009). *Sportmedizin (5. Aufl.)*. Stuttgart: Schattauer.
- IBM Corp (2017). *IBM SPSS Statistics for Windows (Version 25.0) [Computer software]*. Armonk: IBM Corp.
- Jekauc, D., Völkle, M., Wagner, M. O., Mess, F., Reiner, M., & Renner, B. (2015). Prediction of attendance at fitness center: a comparison between the theory of planned behavior, the social cognitive theory, and the physical activity maintenance theory. *Frontiers in Psychology*, 6, 121. <https://doi.org/10.3389/fpsyg.2015.00121>.
- Kerner, C., & Goodyear, V. A. (2017). The motivational impact of wearable healthy lifestyle technologies: a self-determination perspective on fitbits with adolescents. *American Journal of Health Education*, 48(5), 287–297. <https://doi.org/10.1080/19325037.2017.11343161>.
- Kolotylova, T., Koschke, M., Bär, K.-J., Ebner-Priemer, U., Kleindienst, N., Bohus, M., & Schmah, C. (2010). Entwicklung des „Mannheimer Multikomponenten-Stress-Test“ (MMST) [Development of the “Mannheim Multicomponent Stress Test” (MMST)]. *Psychotherapie, Psychosomatik, medizinische Psychologie*, 60(2), 64–72. <https://doi.org/10.1055/s-0028-1103297>.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>.
- Kroidl, R., Schwarz, S., Lehnigk, B., & Fritsch, J. (2015). *Kursbuch Spiroergometrie: Technik und Befundung verständlich gemacht (3. Aufl.)*. Stuttgart: Thieme. <https://doi.org/10.1055/b-003-104201>.
- Krouwer, J. S. (2008). Why Bland-Altman plots should use X, not (Y+X)/2 when X is a reference method. *Statistics in Medicine*, 27(5), 778–780. <https://doi.org/10.1002/sim.3086>.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159. <https://doi.org/10.2307/2529310>.
- Lang, F. R., Martin, M., & Pinquart, M. (2012). *Entwicklungspsychologie – Erwachsenenalter (1. Aufl.)*. Göttingen: Hogrefe. <http://elibrary.hogrefe.de/9783840921865/U1>
- Lejuez, C. W., Kahler, C. W., & Brown, R. A. (2003). A modified computer version of the paced auditory serial addition task (PASAT) as a laboratory-based stressor. *The Behavior Therapist*, 26(4), 290–293.
- Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1), 255–268.
- Liu, J., Tang, W., Chen, G., Lu, Y., Feng, C., & Tu, X. M. (2016). Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai Archives of Psychiatry*, 28(2), 115–120. <https://doi.org/10.11919/j.issn.1002-0829.216045>.
- Lutter, T., Meinecke, C., & Prof, T. (2017). *Zukunft der Consumer Technology – 2017*. <https://www.bitkom.org/sites/default/files/file/import/170901-CT-Studie-online.pdf>. Zugegriffen: 21. Okt. 2020.
- McBride, G. B. (2005). *A proposal for strength-of-agreement criteria for lins concordance correlation coefficient*. NIWA Client Report: HAM2005-062. <https://www.medcalc.org/download/pdf/McBride2005.pdf>
- Mercer, K., Li, M., Giangregorio, L., Burns, C., & Grindrod, K. (2016). Behavior change techniques present in wearable activity trackers: a critical analysis. *JMIR MHealth and UHealth*, 4(2), e40. <https://doi.org/10.2196/mhealth.4461>.
- Meyer, T. (2007). Belastungsuntersuchungen: Praktische Durchführung und Interpretation. In W. Kindermann, H.-H. Dickhuth, A. Nieß, K. Röcker & A. Urhausen (Hrsg.), *Sportkardiologie. Körperliche Aktivität bei Herzerkrankungen (S. 39–66)*. Dordrecht: Springer.
- Meyer, T., Georg, T., Becker, C., & Kindermann, W. (2001). Reliability of gas exchange measurements from two different spiroergometry systems. *International Journal of Sports Medicine*, 22(8), 593–597. <https://doi.org/10.1055/s-2001-18523>.
- Microsoft Corporation (2016). *Microsoft Excel [Computersoftware]*. Verfügbar unter <https://office.microsoft.com/excel>. Zugegriffen: 21. Okt. 2020.
- Nigg, C. R., Borrelli, B., Maddock, J., & Dishman, R. K. (2008). A theory of physical activity maintenance. *Applied Psychology*, 57(4), 544–560. <https://doi.org/10.1111/j.1464-0597.2008.00343.x>.
- Patel, S., Park, H., Bonato, P., Chan, L., & Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of Neuroengineering and Rehabilitation*, 9, 21. <https://doi.org/10.1186/1743-0003-9-21>.
- Peake, J. M., Kerr, G., & Sullivan, J. P. (2018). A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Frontiers in Physiology*, 9, 743. <https://doi.org/10.3389/fphys.2018.00743>.
- Passler, S., Bohrer, J., Blöching, L., & Senner, V. (2019). Validity of wrist-worn activity trackers for estimating VO<sub>2max</sub> and energy expenditure. *International Journal of Environmental Research and Public Health*, 16(17), 3037. <https://doi.org/10.3390/ijerph16173037>
- Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London*, 187, 253–318. <https://doi.org/10.1098/rsta.1896.0007>.
- Pratt, M., Sarmiento, O. L., Montes, F., Ogilvie, D., Marcus, B. H., Perez, L. G., & Brownson, R. C. (2012). The implications of megatrends in information and communication technology and transportation for changes in global physical activity. *Lancet*, 380(9838), 282–293. [https://doi.org/10.1016/S0140-6736\(12\)60736-3](https://doi.org/10.1016/S0140-6736(12)60736-3).
- Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: measures of agreement. *Perspectives in Clinical Research*, 8(4), 187–191. [https://doi.org/10.4103/picr.PICR\\_123\\_17](https://doi.org/10.4103/picr.PICR_123_17).
- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kol-mogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), 21–33. [https://www.researchgate.net/profile/Bee\\_Yap/](https://www.researchgate.net/profile/Bee_Yap/)

[publication/267205556\\_Power\\_Comparisons\\_of\\_Shapiro-Wilk\\_Kolmogorov-Smirnov\\_Lilliefors\\_and\\_Anderson-Darling\\_Tests/links/5477245b0cf29afed61446e1/Power-Comparisons-of-Shapiro-Wilk-Kolmogorov-Smirnov-Lilliefors-and-Anderson-Darling-Tests.pdf](https://www.researchprotocols.org/2018/1/e267205556_Power_Comparisons_of_Shapiro-Wilk_Kolmogorov-Smirnov_Lilliefors_and_Anderson-Darling_Tests/links/5477245b0cf29afed61446e1/Power-Comparisons-of-Shapiro-Wilk-Kolmogorov-Smirnov-Lilliefors-and-Anderson-Darling-Tests.pdf).

- Riebe, D., Ehrman, J. K., Liguori, G., & Magal, M. (Hrsg.). (2018). *ACSM's guidelines for exercise testing and prescription* (10. Aufl.). Philadelphia, Baltimore, New York: Wolters Kluwer.
- Scharhag-Rosenberger, F., & Schommer, K. (2013). Die Spiroergometrie in der Sportmedizin. *Deutsche Zeitschrift Für Sportmedizin*. <https://doi.org/10.5960/dzsm.2013.105>.
- Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5, 258. <https://doi.org/10.3389/fpubh.2017.00258>.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591–611. <https://doi.org/10.2307/2333709>.
- Statista Digital Market Outlook (2020). *Fitness report 2019: statista digital market outlook – segment report*. <https://de.statista.com/statistik/studie/id/36673/dokument/fitness-outlook/>. Zugegriffen: 19. Okt. 2020.
- Tanosoto, T., Bendixen, K. H., Arima, T., Hansen, J., Terkelsen, A. J., & Svensson, P. (2015). Effects of the Paced Auditory Serial Addition Task (PASAT) with different rates on autonomic nervous system responses and self-reported levels of stress. *Journal of Oral Rehabilitation*, 42(5), 378–385.
- Tarvainen, M. P., Niskanen, J.-P., Lipponen, J. A., Ranta-Aho, P. O., & Karjalainen, P. A. (2014). Kubios HRV—heart rate variability analysis software. *Computer Methods and Programs in Biomedicine*, 113(1), 210–220. <https://doi.org/10.1016/j.cmpb.2013.07.024>.
- Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996). Heart rate variability. *European Heart Journal*, 17(3), 381–385. [https://doi.org/10.1093/eurheartj/17.suppl\\_3.381](https://doi.org/10.1093/eurheartj/17.suppl_3.381).
- The Cooper Institute® (2007). *Physical Fitness Assessments and Norms for Adults and Law Enforcement*. The Cooper Institute: Dallas, Tx.
- Watson, P. F., & Petrie, A. (2010). Method agreement analysis: a review of correct methodology. *Theriogenology*, 73(9), 1167–1179. <https://doi.org/10.1016/j.theriogenology.2010.01.003>.