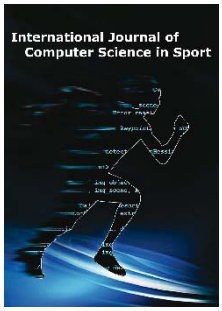
International Journal of Computer Science in Sport

Volume 18, Issue 3, 2019

Journal homepage: http://iacss.org/index.php?id=30

DOI: 10.2478/ijcss-2019-0016



Evaluation of mobile applications for fitness training and physical activity in healthy low-trained people - A modular interdisciplinary framework

Josef Wiemeyer

Institute for Sport Science, Technische Universität Darmstadt, Germany

Abstract

Numerous mobile applications are available that aim at supporting sustainable physical activity and fitness training in sedentary or low-trained healthy people. However, the evaluation of the quality of these applications often suffers from severe shortcomings such as reduction to selective aspects, lack of theory or suboptimal methods. What is still missing, is a framework that integrates the insights of the relevant scientific disciplines.

In this paper, we propose an integrative framework comprising four modules: training, behavior change techniques, sensors and technology, and evaluation of effects. This framework allows to integrate insights from training science, exercise physiology, social psychology, computer science, and civil engineering as well as methodology. Furthermore, the framework can be flexibly adapted to the specific features of the mobile applications, e.g., regarding training goals and training methods or the relevant behavior change techniques as well as formative or summative evaluation.

KEY WORDS: MOBILE TECHNOLOGIES, WEARABLES, FITNESS TRACKER, EVALUATION FRAMEWORK, SUSTAINABLE PHYSICAL ACTIVITY

Introduction

Sustainable physical activity (SPA) and fitness training (FT) have become an indispensable and constitutive part of health-related activities (e.g., ACSM, 2011; WHO, 2010, 2018). SPA and FT have numerous benefits for health, ranging from positive effects on physiological, e.g., cardiovascular, respiratory, metabolic, neuromuscular and hormonal, functions and anatomical structures of the human organism, e.g., heart, blood vessels, blood, muscles, bones, and brain, to positive effects on cognition, emotion, volition, and motivation in all age groups (Janssen & LeBlanc, 2010; Poitras et al., 2016; WHO, 2018). Furthermore, SPA and FT support prevention and therapy of numerous diseases, e.g., heart diseases, stroke, diabetes mellitus and specific types of cancer (WHO, 2018). Despite these tremendous benefits, many people do not meet the minimum requirements for SPA. "Worldwide, 1 in 4 adults, and 3 in 4 adolescents (aged 11–17 years), do not currently meet the global recommendations for physical activity set by WHO" (WHO, 2018, p.6). Furthermore, the initiation and maintenance of regular physical activity (PA) and FT poses numerous challenges to many people. The list of impediments and barriers ranges from lack of motivation or attitude over fear of injury or falling to time and financial constraints (e.g., Lachman et al., 2018).

Strictly speaking, the ultimate goals must be to initiate and maintain a change of behavior in humans that have previously been inactive or did not meet the required amount or continuity of SPA yet. To establish sustainable engagement in PA, social psychology claims that people (are able to) adopt behavior change techniques (BCT; e.g., Williams & French, 2011; Michie et al., 2011, 2013). These techniques include various behavioral, (meta-)cognitive, social, emotional, volitional, and motivational mechanisms of support that are derived from several theories.

Beyond establishing SPA, a high quality of training must be offered, including instructions and plans for training, adequate monitoring and feedback etc. Physical training is at least ineffective or inefficient if not harmful if the current insights from sport science, particularly training and movement science as well as sport medicine and exercise physiology, are ignored (e.g., Halson et al., 2016).

In the age of ubiquitous information and communication technologies (ICT), the question arises whether and how SPA and PT in healthy sedentary or low-engaged target groups can be supported or enhanced by the usage of mobile and wearable applications. Considering the claims of the publishers, mobile fitness apps promise to boost fitness or health training. However, these "aggressive and exaggerated claims" (Düking et al., 2018, p.1) have rarely been sufficiently validated (e.g., Halson et al., 2016; Peake, Kerr, & Sullivan, 2018; Romeo et al., 2019). In current reviews, two aspects are criticized, i.e., lack of user or consumer integration in the development process and lack of adequate validating research (e.g., Düking et al., 2018; Peake, Kerr, & Sullivan, 2018, Warraich, 2016). Halson et al. (2016) enumerate numerous issues of mobile ICT regarding SPA and FT, ranging from technical issues like sensor placement and accuracy (see also Wahl et al., 2017) to ethical considerations. Furthermore, motivational and informational challenges have to be met (Schmidt et al., 2015).

Therefore, the goal of this paper is to present an appropriate framework for the evaluation of fitness apps. This framework is intended to assess the quality of the fitness app for SPA and FT. Regarding fitness apps, "quality" can pertain to different dimensions: outcomes, application procedures and conditions, and features of the technical system.

Considering the aspects discussed earlier, from a scientific point of view, the following criteria constitute the quality of fitness apps:

1. *Outcomes*: The fitness app must be effective and efficient. This means that the app has to accomplish the respective training goals with an adequate amount of investment of time, money etc. Effectivity and efficiency should be evaluated using appropriate scientific approaches to formative and summative evaluation.
2. *Application procedures (training)*: In order to reach the training goals, the app has to employ appropriate scientific concepts and rules for training.
3. *Technical system*: Within the respective training framework, the app has to deliver the required information with adequate accuracy and precision. This means that technology and in particular sensors have to work accurately and precisely.
4. *Application procedures (control of behavior)*: Finally, training aims at sustainable engagement in PA. This means that the respective scientific models of change and maintenance of behavior have to be included in the app.

Therefore, four modules are included in the framework for evaluating the quality of mobile fitness apps: Training (T), Behavior change techniques (B), technology (T) and evaluation of effects (E). Note that this framework does not cover all aspects of training with mobile apps. As has been argued previously, further aspects such as legal and ethical issues (e.g., privacy policy) are also important factors contributing to the quality of mobile fitness apps (e.g., Bondaronek et al., 2018). However, regarding the primary goal of sustainably improving the fitness and PA of healthy people, these aspects are considered secondary.

In this paper, we discuss the four modules of the TBTE framework: (1) the sport-scientific basis of PA and methods for FT, (2) the social-psychological basis of BCT, (3) mobile and wearable technologies, and (4) evaluation of the effects of mobile ICT on SPA and FT. In each section, we start with the scientific basics of the respective field, followed by selected examples, studies and reviews addressing the application of mobile ICT in the respective field. Finally, we discuss the complete TBTE framework for assessment and evaluation of mobile apps for SPA and FT.

Considering the wide range of physical training and due to the fact that the primary target groups of mobile fitness apps are healthy people with a low engagement in PA and low fitness level this paper focusses on the group of healthy untrained or low or moderately trained persons. Another reason for excluding other target groups is that the requirements and conditions for high- or top-level athletes as well as people suffering from diseases are very specific and much more complex compared to lower levels of training in healthy people.

Structure of PA, physical fitness and methods for physical exercise

Physical Activity (PA) is defined as “any bodily movement produced by skeletal muscles that requires energy expenditure” (WHO, 2010, p.53). This means, that human movements have to exceed a certain threshold (i.e., basic metabolic rate) to be considered PA. This is established by activating the big muscles of the body, e.g., leg, arm or trunk muscles. Typical examples are walking, running, rowing, and cycling. Energy expenditure (EE) can be measured in units of Calories or Joule. An alternative is to determine MET, i.e., metabolic equivalent of task (Ainsworth et al., 1993). For example, a value of 7 MET (moderate cycling or jogging) means, that a person with a weight of 70kg spends 490 kcal per hour in the respective PA. PA can be classified as low or light (EE < 3 MET), moderate (EE between 3 and 6 METs) and vigorous (EE > 6 MET). Beyond intensity, PA can also be quantified according to duration, volume, frequency, and density (i.e., relation of load periods and breaks). In addition to increase of EE,

PA is accompanied by or results in several biomechanical, physiological, and psychic processes (see also Table 3), for example, force production, acceleration of body and body parts, increase of heart rate (HR), and increase of rating of perceived exertion (RPE).

Physical fitness (PF) is defined as "a set of attributes people have or achieve that relate to the ability to perform physical activity" (Caspersen, Powell, & Christenson, 1985, p.129). PF can be categorized into energy-determined attributes such as strength, endurance, and speed and information-determined attributes such as sensorimotor coordination and skills. Important health-related components of PF are cardiorespiratory, neuromuscular, metabolic, and sensorimotor functions such as balance, agility, and reaction (ACSM, 2011).

Physical training, exercise or fitness training (FT) can be defined as "subset of physical activity that is planned, structured, and repetitive and has as a final or an intermediate objective the improvement or maintenance of physical fitness" (Caspersen, Powell, & Christenson, 1985, p.126). FT recommended for the primary target group of fitness apps by international institutions (e.g., ACSM, 2011; WHO, 2010) includes four main components of physical fitness:

- Aerobic capacity
- Strength or resistance
- Flexibility
- Sensorimotor coordination and skills

Regarding these four components, specific training methods exist. These methods can be generally characterized according to the FITT-VP framework (Gibson, Wagner, & Heyward, 2018, p.126-127), which is well-accepted in training science.

- Frequency, i.e., number of sessions per week
- Intensity, i.e., level of training stimulus
- Time, i.e., duration of training stimulus or number of sets/repetitions or density of training stimulus (relation of load and recovery phases)
- Type of training stimulus
- Volume, i.e. total duration or total training load
- Progression

In addition, further indicators of training load have been proposed, for example, fractional and temporal distribution of the contraction modes (static, concentric, eccentric) per repetition, duration of one repetition, rest in-between repetitions, time under tension, muscular failure, range of motion, recovery time, and anatomical definition for resistance training (Toigo & Boutellier, 2006; Wackerhage et al., 2018) or the SPORT approach (specificity, progression, informational overload, reversibility, and tedium) for skill training (Farrow & Robertson, 2017).

Existing recommendations for training in the respective target group (e.g., ACSM, 2011; O'Donovan et al., 2010; WHO, 2010) are mainly based on the FITT-VP framework. In Table 1, these recommendations are summarized.

Table 1. Recommendations for PA and FT in healthy untrained or moderately trained people according to ACSM (2011), WHO (2010) and O'Donovan et al. (2010)

Type	**Intensity**	**Time/ Density**	**Frequency [d/wk]**	**Volume**	**Further recommendations**
Aerobic endurance	Moderate (3-6 MET) or Vigorous (> 6 MET)	≥ 10 min (single bout) continuous	≥ 5	150 min/wk ≥ 500 – 1000 MET min/wk 800 – 1,000 kcal/wk	Cyclic movements, major muscles P: gradual - volume
Strength	40-70% 1RM 8 – 12RM	Discontinuous break: 2 – 3 min between sets	2 - 3	1 – 3 sets 8 – 15 reps. 30 min/wk	Major muscle groups 8 – 12 exercises P: gradual - intensity, reps, frequency
Flexibility	Feeling of tightness/ slight discomfort	10 – 60 sec	2 - 3	60 sec per exercise	Active or passive static stretch, dynamic stretch or PNF stretch
Sensorimotor control	Informational (over)load	≥ 20 – 30 min/d	2 - 3	≥ 60 min/wk	Motor skills (sport, leisure, ADL) & abilities (balance, agility, coordination) Proprioceptive exercises P: difficulty, complexity, reps

Legend: reps – repetitions; RM – repetition maximum (i.e., load the allows for the respective reps); MET – metabolic equivalent of task; P – progression; ADL – activities of daily living

Table 1 shows that in order to achieve specific adaptations, the training stimulus has to be specific and exceed a certain threshold. To be effective, training load has to be tailored to the individual conditions such as fitness level, health conditions, and age (e.g., O'Donovan et al., 2010; WHO, 2010). Finally, short-term and long-term recovery play an important role (e.g., Kellmann et al., 2018).

In addition to the general recommendations, specific training methods exist that are also relevant to training in healthy low-trained people, e.g., (high-intensity) interval training for time-efficient aerobic exercises comprising short bouts of near-maximum intensity with short, incomplete breaks (e.g., Batacan et al., 2017) or PNF (proprioceptive neuromuscular facilitation) stretching methods including pre-stretch isometric contraction of the agonists and peri-stretch contraction

of antagonists for more effective increase of range of motion (e.g., Guissard, Duchateau, & Hainaut, 1988; Fasen et al., 2009).

The methods of training serve as components for a training plan. This training plan contains several elements:

- *Training goals*: anticipated and desired effects of training
- *Training content*: methods, media and exercises (including training devices)
- *Training schedule*: temporal distribution of training sessions dedicated to specific training goals and content

Furthermore, training has to be systematically monitored to ensure accomplishing the training goals, to prevent adverse events such as injury, illness, overreaching and overtraining and to support sustainable exercise motivation. This iterative process is illustrated in Figure 1. Training is executed according to the training plan. This execution is documented by training protocols, athlete monitoring (e.g., diaries or logs), and performance diagnostics (e.g., field or laboratory tests). These documents and data are analysed regarding process and outcomes according to the training plan. It may be necessary to correct and modify the training plan, if the analysis shows significant discrepancies, e.g., between intended and actual outcomes.

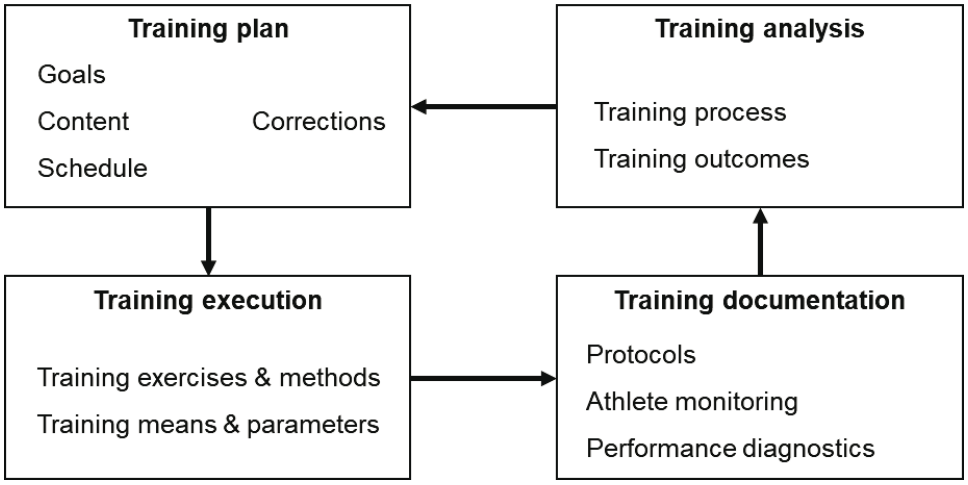


Figure 1. Iterative process of training control (modified according to Hohmann, Lames & Letzelter, 2002, p.167)

Based on the recommendations of health institutions (ACSM and NASM), Chi-Wai et al. (2011) developed a Virtual Fitness Training Workflow (VFTW) including four main stages: pre-participation assessment (5 steps), pre-exercise evaluation and exercise prescription (8 steps), program monitoring (7 steps), and program evaluation (4 steps). Particularly, the first stage and parts of the second stage expand the model illustrated in Figure 1 by specifying information required for building an individualized training plan, i.e., assessment of trainees' expectations, preferences and constraints, health status and medical history, and performance level.

To conclude this section, training science, exercise physiology and sports medicine provide numerous important insights into the process of PA and FT that have to be considered in mobile training apps in order to ensure a sound basis of successful training. A checklist for evaluating mobile apps for PA and FT should at least contain the following domains and functions (Chi-Wai et al., 2011; Kranz et al., 2013; Wiemeyer et al., 2016; Kettunen, Critchley & Kari, 2019):

- *Assessment and input of individual characteristics*, e.g., age, gender, anthropometrics (body height and weight), performance level, training goals, health status, expectations, preferences, and constraints

- Creation and input of *individualized and flexible training plans* according to the characteristics mentioned above
- Data base including a *collection of exercises* for FT, indexed by appropriate meta-data
- Data base including *tests and procedures for diagnostics* of performance and fitness level as well as health status, preferences, and constraints
- *Recording and feedback functions* regarding quality of training, including training load (stress), objective and subjective strain indicators (e.g., Borg scale) and performance indicators (adaptation)
- Options for the *analysis* and (audio-visual) *presentation* of the recorded training data
- Online and offline *feedback functions* to inform and motivate the trainees
- Prescriptions, guidelines and recommendations for *individualized training*
- *Coaching functions* including social support and persuasive strategies, context-related instructions, goal-oriented, individualized and immediate feedback, motivation, appraisal and reinforcement, and advice

Behaviour change techniques (BCT)

Beyond requirements derived from training science, sustainable FT and PA means to adopt and maintain a change of behaviour. Because this issue is not specific to SPA and FT, numerous social-psychological models regarding sustainable change of behaviour have been proposed. These models can be categorized according to specificity (generic versus domain-specific) and temporal criteria (structural or static versus procedural or dynamic; Table 2). Whereas generic models address human behaviour in general, specific models focus on PA. Structural models describe the interaction of more or less time-invariant factors whereas dynamic models focus on temporal sequences of particular phases. The models address selected cognitive, social-cognitive, emotional, motivational and volitional determinants of human behaviour.

Important **generic structural models** that have been successfully applied to PA are the theory of planned behaviour (TPB; Ajzen, 1991; Hagger, & Chatzisarantis, 2014), the social-cognitive theory (SCT; Bandura, 1999), and the self-determination theory (SDT; Ryan & Deci, 2000; Teixeira et al., 2012). According to the TPB, three factors indirectly contribute to behaviour via intention: perceived control, subjective norm, and attitudes towards the behaviour. Hagger and Chatzisarantis (2014) added implicit attitudes and implicit motivation as factors directly influencing behaviour. The SDT claims that three important factors influence intrinsic motivation, i.e., autonomy, competency, and social relatedness. Going beyond the individuum, the SCT claims that human behavior emerges from a dynamic and flexible interaction of internal personal factors, behavioral patterns, and environmental influences. One important assumption of this theory is the distinction between direct personal agency, proxy agency (i.e., delegating agency to other persons), and collective agency (i.e., acting in groups).

Generic procedural models include the trans-theoretical model (TTM; Prochaska, Redding, & Evers, 2008; Marshall & Biddle, 2001) and the Rubicon model (Heckhausen, 1989). Whereas the TTM postulates six stages in the long-term adoption of new behaviour, i.e., precontemplation, contemplation, preparation, action, maintenance, and termination, the Rubicon model distinguishes two volitional (i.e., action initiation and maintenance) and two motivational phases (i.e., formation and deactivation of intentions) in short-term control of goal-directed behaviour. The TTM includes numerous cognitive, motivational, and social processes happening in the six stages.

Domain-specific structural models of PA claim a more or less time-invariant interplay of factors contributing to sustainable change of behaviour. For example, the Health Belief Model (HBM; Champion & Skinner, 2008) claims an important influence of individual beliefs regarding susceptibility of disease, threats, benefits, barriers, and self-efficacy on health-related behaviour. Another structural model by Wagner (2000) confirmed a medium term (6 months) influence of group affiliation, social support by family and health satisfaction on maintenance of PA as well as a long-term (12 months) influence of self-efficacy, social support, and intention.

A well-known and confirmed **domain-specific process model of health-related behaviour** is the MoVo concept integrating motivational and volitional factors influencing human behaviour change (Fuchs et al., 2011). This model claims that intentions and situational cues directly and indirectly influence the initiation of actions. Furthermore, outcome experiences, outcome expectancies, and self-efficacy indirectly influence the initiation and maintenance of actions, particularly in early stages. Finally, the (delayed) influence of self-concordance (i.e., extent to which a specific goal intention is in accordance with the attitudes of the person) and barrier management could be confirmed (Fuchs et al., 2012).

Table 2. Examples of models relevant to sustainable adoption of training

Reference / **Specificity**	**Time (procedural)**	**Structure**
Generic	Trans-Theoretical Model	Theory of Planned Behaviour
	Rubicon model	Social-Cognitive Theory
		Self-Determination Theory
Domain-specific	MoVo model	Health Belief Model

Beyond these models, various authors have collected and classified BCT as a "union" (or taxonomy) of different theories. Munson and Consolvo (2012) identified four "promising approaches" for BCT: goal-setting, rewards, self-monitoring, and sharing data and experiences. Michie et al. (2013) present a taxonomy of 93 BCT which have been ordered in 16 clusters: scheduled consequences, rewards and threats, repetition and substitution, antecedents, associations, covert learning, natural consequences, feedback and monitoring, goals and planning, social support, comparison of behaviour, self-belief, comparison of outcome, identity, shaping knowledge, and regulation. The BCT are located at the (meta-)cognitive, motivational, emotional and social level.

Regarding technology-based interventions aiming at weight loss, Khaylis et al. (2010) discovered five key factors: self-monitoring (e.g., diary or activity recording), counsellor feedback and communication (e.g., motivating feedback regarding goals, results, and progress), social support (e.g., chats with peers and friends), use of a structured program (e.g., regular lessons), and use of an individually tailored program (e.g., tailored to individual goals, preferences, and barriers). Rhea, Felsberg and Maher (2018) proposed a theory-based framework to support developing of health apps based on scientific evidence.

Considering this huge amount of BCT, the question arises whether all BCT are equally important for SPA and FT. Williams and French (2011) meta-analysed 27 PA-directed intervention studies. They found that the following BCT had a significantly low to moderate impact on PA: facilitation of social comparison (Effect size d=0.46), action planning (d=0.38), reinforcement

of effort or progress (*d*=0.33), time management (*d*=0.33), provision of information on consequences of behaviour in general (*d*=0.27), and provision of instruction (*d*=0.26). Regarding self-efficacy, an important long-term source of maintenance, action planning (*d*=49), social comparison (*d*=0.34), reinforcement (*d*=0.31), and instruction (*d*=0.21) had a low to moderate influence.

BCT have been applied in several studies evaluating mobile PA apps.

Conroy, Yang and Maher (2014) assessed 167 top-ranked mobile PA apps based on the CALO-RE taxonomy (Michie et al., 2011). They found that most of the apps incorporated fewer than four BCT; the five most common BCT were: "instruction on how to perform exercises, modelling how to perform exercises, providing feedback on performance, goal-setting for physical activity, and planning social support/change" (Conroy, Young & Maher, 2014, p.649).

Direito et al. (2014) analysed the top-20 payed and top-20 free apps for PA and nutrition using a 26-BCT taxonomy (Abraham & Michie, 2008). They found that paid apps contained more BCT (M = 9.7; *range* 2 – 18) than free apps (M = 6.6; *range* 3 – 14). More than half of the apps addressed provision of instruction, graded task-setting, prompting self-monitoring and identification as a role model, planning of social support/change, social comparison, and feedback on performance. Prompts were used more extensively in paid apps, while demonstration of behaviour and provision of information on consequences were more present in free apps.

C.H. Yang, Maher and Conroy (2015) evaluated 100 PA apps using the 93-BCT taxonomy of Michie et al. (2013). An average of 6.6 BCT (*Mdn*=6) was incorporated in each app. About 50% of the apps included social support (79%), information about others' approval (64%), instruction on how to perform a behaviour (49%) and demonstration of the behaviour (47%), followed by feedback on behaviour (42%) and goal setting (36%).

McKay et al. (2018) reviewed 36 studies evaluating apps dedicated to health issues. From the 5 studies related to PA (quality: 9 to 13 of 15 points; medium to high), three (i.e., the studies analysed above) applied checklists based on theory, and two studies (reported in the fourth section) assessed effectiveness of PA apps.

Shameli et al. (2017) analysed the data of 3,637 users of a competitive PA app. The users were engaged in 2,432 seven-day competitions. The authors found a significant effect of social competition on PA (measure: daily steps) which was independent of gender, age, and baseline activity. An important factor influencing engagement is the closeness of competition; in close competitions, engagement is much higher compared to competitions with big differences between the competitors. The most important single predictor of PA engagement was the engagement in previous competitions.

McKay, Slykerman and Dunn (2019) have recently proposed an "App Behavior Change Scale" (ABACUS) for assessing the BCT quality of mobile applications. The ABACUS comprises 21 items constituting 4 scales: Knowledge and information (5 items: options for customization/personalization, informed development, input of baseline information, instruction, consequences of behaviour), goals and planning (3 items: readiness, goal setting, goal monitoring), feedback and monitoring (7 items: understandable feedback, self-monitoring options, social comparison, automatic or personal feedback, data export options, rewards and incentives, encouragement and reinforcement), and actions (6 items: reminders, prompts, and cues, encouragement of positive habit formation, no limits for exercise, plan for barriers, support for restructuring environment, support with distraction and avoidance). Unfortunately, only 10 items reached sufficient interrater reliability (Krippendorff $\alpha \geq .5$) in the final evaluation,

whereas overall reliability (ICC = .91) and internal consistency were high (Cronbach $\alpha = 93$).

To conclude this section, the following checklist for evaluating mobile apps for PA and FT regarding BCT is reasonable:

- *Action planning*: goals (set and monitor), consequences, barriers, alternatives, schedule
- *Feedback, reinforcement and rewards*: effort, outcomes and progress as well as causal attribution (intrinsic – changeable causes)
- *Meta-cognitive strategies*: resource management like effort, time and barriers, self-monitoring
- *Informational guidance*: Knowledge, information, feedback and instruction
- *Support options* for prompts, cues, and reminders
- *Social comparisons and competition*
- *Social support*: family, friends, training group (for different types of social support, see Chi-Wai et al., 2011, p.57)
- *Supporting intrinsic motivation, self-efficacy and self-concordance*, e.g. by individualization and customization, role models, feedback, and appropriate causal attribution

Mobile technologies

„Technological development has promoted the emergence of various new technologies that allow their user to track, measure, and evaluate a multitude of personal activities and biosignals" (Kari & Rinne, 2018, p.128). Heikenfeld et al. (2018, p.217) even call it an "explosion of wearable sensors".

PA and FT include movements of body parts and the whole body. These movements can be assessed and analysed regarding three dimensions, i.e., a biomechanical, a physiological, and a psychic dimension including numerous parameters (see Table 3).

The options for assessment of PA and FT documented in Table 3 are currently not fully deployed in mobile apps dedicated to the target group. In the majority of applications, standard smartphone sensors (i.e., accelerometer, gyroscope, orientation sensors, camera, and GPS, He & Li, 2013) are used to assess simple PA indicators such as step or activity counts, covered distance, velocity or activity (profiles), total training time, repetitions and energy expenditure (e.g., Knight et al., 2015; Wiemeyer et al., 2016). In addition, smartphones can be coupled to a bunch of wireless biosensors such as optical or electric sensors via USB, Bluetooth, ANT, ZigBee or WiFi, to measure HR (Ludwig et al., 2018), hormones, electrolytes, or metabolites (Roda et al., 2016; Kassal, Steinberg, & Steinberg, 2018). Currently, the most commonly measured parameters from external sensors are HR and EE (Knight et al., 2015). For the future, considering the further development of smartphone and sensor technology as well as appropriate algorithms for signal processing and classification, these potentials are expected to be more extensively exploited.

Table 3. Overview of dimensions of PA and FT and options for assessment

Dimension	**Parameter** **Examples**	**Assessment (sensors, scales)**	**Reference**
Biomechanical	Kinematics: Velocity, acceleration Joint position & angle	Accelerometer Inertial sensors Goniometer GPS	Baca (2015) Mukhopadhyay (2015) Heikenfeld et al. (2018)
	Kinetics: Force, torque Work, energy, power	Resistive sensors Piezoelectric sensor Capacitive sensor	
Physiological	Cardiovascular: HR, HRV, ABP, ECG	numerous sensors and principles, e.g., electric or optical methods	Baca (2015) Ludwig et al. (2018)
	Respiratory: Respiratory rate, VT; VO_2	Spirometric sensors	Baca (2015) Heikenfeld et al. (2018)
	Metabolic: Lactate, glucose; Energy expenditure	Invasive and non-invasive sensors (e.g., amperometric)	Gao et al. (2016); Roda et al. (2016)
	Hormones: Cortisol, testosterone	Invasive and non-invasive sensors	Heikenfeld et al. (2018) Roda et al. (2016)
	Electrolytes: sodium (Na^+), potassium (K^+)	Invasive and non-invasive sensors (ion-selective membrane cocktails)	Gao et al. (2016)
	Neuromuscular: EMG	EMG surface electrodes	Baca (2015) Wong et al. (2015)
Psychic	Subjective strain: Exertion	RPE (Borg scales)	Borg (1998)
	Recovery	Recovery-stress questionnaire	Kellmann & Kallus (2001) Kellmann et al. (2018)

Dimension	Parameter Examples	Assessment (sensors, scales)	Reference
	Emotion:	POMS	Leunes & Burger (2000)
	Mood, enjoyment	PANAS	Watson, Clark & Tellegen (1988)
		PACES	Kendzierski & DeCarlo (1991).
	motivation	SMS	Pelletier et al. (1995)

Legend: GPS – global positioning system; ECG – Electrocardiogram; HR – heart rate; HRV – HR variability; ABP – arterial blood pressure; VT – ventilatory threshold; VO_2 – oxygen uptake; EMG – electromyography; RPE – rating of perceived exertion; POMS – profile of mood states; PANAS – positive affect negative affect scale; PACES – physical activity enjoyment scale; SMS – sport motivation scale

Regarding the digital acquisition of physical signals, there is a chain of processes that comprise signal transduction, conditioning, A/D conversion, transmission, and further digital processing (see Figure 2). The primary criteria for quality of this chain is errors of measurement, i.e., consistency and deviation of measurement from the "true" value (reliability and validity; e.g., Atkinson & Nevill, 1998). Errors can be divided into systematic error or systematic bias (accuracy) and random error (precision). Furthermore, static and dynamic error can be distinguished. The sources of errors are manifold, ranging from sensor inaccuracy, noise in the system, inadequate representation to calculation error or human mistakes. For example, a bioelectrical HR sensor may be inadequately located at the breast or wrist and may therefore lose skin contact or wireless signal transmission may be interrupted. In addition, processing an erroneous HR signal to calculate EE or a noisy acceleration signal to calculate distance may cause error propagation problems.

There are numerous statistical measures of error, ranging from simple descriptive measures such as various mean difference measures (e.g., mean difference, mean absolute difference error – MAPE - or root mean square error – RMSE) and coefficient of variance (CV) to inferential statistics such as hypothesis testing or correlation, regression and structural equations (see Table 4). In addition, the graphical representation of error assessment also covers a wide range from simple bar charts to regression and Bland-Altman plots (Atkinson & Nevill, 1998).

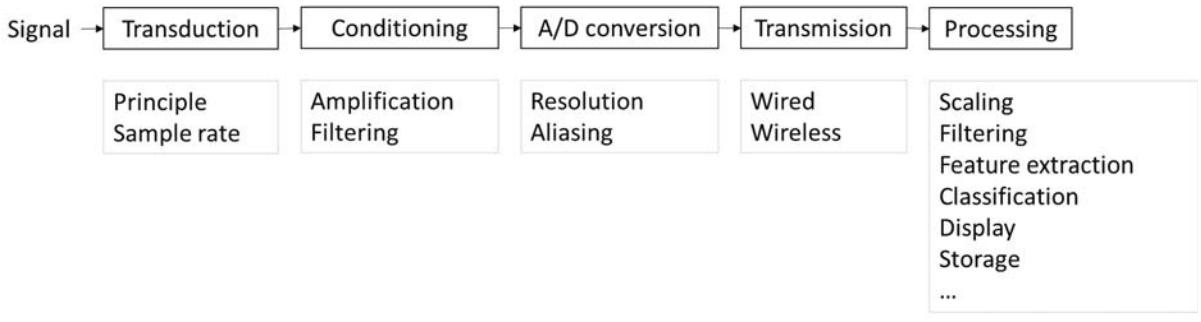


Figure 2. Chain of digital data acquisition and signal processing

Table 4. Selected statistics for error assessment (taken from Atkinson & Nevill, 1998)

Measures	Remarks
t-Test, repeated measures ANOVA	Depending on random error, no post measurement correction possible
	Depending on sample homogeneity
Pearson correlation	Different ways of calculation
ICC	Inappropriate statistical model
Regression	
SEM	No clarity regarding acceptable SEM, several pre-assumptions
CV	
Limits of agreement	Normality required, no agreed limit
	Bland-Altman plot (visual exploration), systematic and random error

Legend: ANOVA – Analysis of Variance; CV – Coefficient of Variation; ICC – Intraclass correlation; SEM – Standard Error of Mean

Table 5. Recommendations for standardized evaluation of PA-monitoring wearable devices (supplemented according to Düking et al., 2018, p. 4)

Category	Factor	Specification (examples)
Wearable-specific factors	Sensor characteristics	Scrutiny of each sensor
	Software	Calculations, algorithms; version
	Raw data	Sampling frequency and pre-processing (e.g., filtering, amplification)
	Durability	Durability and age of the device
	Anatomical positioning	Report of exact positioning, reproducibility and possible interferences
Evaluation conditions	Study population	Description, inclusion & exclusion criteria
	Exercise protocol	Detailed description; systematic variation of form and intensity of exercise
	Potential confounders	Report, check, control and minimize influence
Statistical analysis	Selection and calculation of adequate measures of …	
	Reliability	Intra- and inter-device reliability
	Sensitivity & selectivity/ specificity	Smallest worthwhile change (continuous data); true positives & true negatives (classification)
	Validity	Concurrent criterion, content, and construct validity

Düking et al. (2018) propose recommendations for standardized evaluation of reliability, sensitivity, and validity of PA-monitoring wearable devices. In their checklist the authors identify 11 factors which are grouped into three categories (see Table 5).

Regarding accuracy (systematic error or bias) and precision (random error) of the sensors integrated in the mobile apps, numerous studies have been published showing a great variety in measures and procedures. For example, Case et al. (2015) compared 10 different applications and devices (4 smartphone apps, 1 pedometer, 2 accelerometers, 3 wearable devices) concerning predetermined number of steps (500 and 1500 steps on a treadmill at 3 mph) in a sample of 28 health adults aged 18 years and above. With one exception, the devices showed reasonable accuracy and precision (measures: absolute and relative systematic error; SD for random error); however, the accelerometers (FitBit One and Zip) were more accurate and precise compared to all other devices. Note that the devices were placed at different body locations, i.e., waistband (accelerometers and pedometer), wrist (wearable devices), or pants pocket (smartphones); therefore, the advantages of the accelerometers may be due to their placement at the waistband, i.e. reduced representation error.

Battenberg et al. (2017) also tested 10 step-measuring devices (9 accelerometer-based apps and one pendulum-design pedometer) in a sample of 20 healthy young adults (mean age: 25.6 years; 12 females and 18 males). The protocol included five exercises (400 m brisk walk or run, 10 m walk at "household pace", ascend 10 steps, and descend 10 steps) and was repeated three times. Again, placement of the sensors varied (waistband, wrist, ankle, and anterior superior iliac spine). Accuracy of the devices was determined by percentage error (formula: 100*(device count – actual count)/actual count) and showed considerable variability depending on the exercise. Overall, one accelerometer-based waist-born device (FitBit One) showed highest accuracy (above 94%) and precision (smallest 95% confidence intervals). The most challenging conditions were slow walk and stair-climbing.

Kooiman et al. (2015) checked the accuracy of 10 fitness trackers using a sample of 33 healthy adults (16 males, 17 females; aged: between 18 and 64 years). The authors tested the devices under laboratory (treadmill: 30-minute walk at 4.8 km/h) and field conditions (normal working day between 9.00 am and 4:30 pm). The OptoGait and ActivPAL system served as reference systems. Accuracy was determined by the difference between tracking device and gold standard (OptoGait treadmill system) and MAPE (formula: 100*(mean difference device –gold standard)/ mean gold standard). Reliability was determined by ICC between device and gold standard, while level of agreement was assessed by Bland-Altman plots. Overall, the FitBit Zip device showed highest accuracy (validity) and reliability.

Fokkema et al. (2017) found that accuracy and precision of 10 step-counting fitness trackers varies, i.e. generally increases (with some exceptions) with running speed (3.2, 4.8, and 6.4 km/h). Accuracy was determined by mean difference between device and gold standard (manual hand counter) and MAPE (cut-off criterion: 5%), while precision was assessed by ICC. Only one tracker (Apple watch) showed good accuracy and precision at all speeds. Test-retest reliability was best in Samsung Gear S and FitBit Charge HR.

Bender et al. (2017) assessed the accuracy (relative differences and correlations between two devices) of four fitness trackers in a 14-week intervention under field conditions (sample: 2 males, 1 female; members of the research team). Participants wore the devices for approximately 16 hours for pairwise comparison on three days. Whereas differences between two devices of the same product (FitBit Flex) were low (range: 0 to 7%), differences between different products (FitBit Charge HR versus Garmin vivoactive; FitBit Flex versus Apple Watch) were high (range: 0 to 40%).

Wahl et al. (2017) tested the criterion validity of 11 different wearables (Bodymedia Sensewear, Beurer AS 80, Polar Loop, Garmin Vivofit, Garmin Vivosmart, Garmin Vivoactive, Garmin Forerunner 920XT, Fitbit Charge, Fitbit Charge HR, Xaomi MiBand, Withings Pulse Ox) in a sample of 20 healthy students. The participants performed a protocol including various running exercises (i.e., staged, intermittent, and outdoor runs). Validity was assessed by MAPE (formula: 100*(mean difference device – gold standard)/ mean gold standard) and ICC (device versus gold standard) as well as typical error (TE; formula: TE = SD · √1-ICC) and upper and lower limits of agreement according to Bland-Altman. Whereas the devices showed good validity regarding step count, measurement of covered distance and energy expenditure was not sufficiently valid.

Adopting a different (qualitative) approach, R. Yang et al. (2015) analysed 600 product reviews and interviewed 24 end-users of six mobile fitness trackers (3 wrist-worn, 2 waist-worn and one chest-worn). The accuracy issue was mentioned in 220 of the 600 reviews. Analysis revealed that users had a different understanding of "accuracy"; most of them rather meant "reliability" or "precision" of measurements. To test the accuracy or precision of their devices, users adopted more or less intuitive procedures, e.g., adhoc or spontaneous assessment and folk-testing as well as comparison with ground truth or other commercial devices. Considering these more or less unsystematic and error-prone procedures, the authors recommend to include testability as a feature into the applications.

The physiological signal measured most often by mobile PA and FT apps is heart rate (HR; Mukhopadhyay, 2015). Regarding HR, numerous procedures and devices are available (Ludwig et al., 2018). Electrographical procedures such as electrocardiogram (ECG) are considered the gold standard; they show the best validity compared to other procedures. For many other procedures, validity is either worse or data are not (yet) available. For example, Ho et al. (2014) compared four smartphone apps with the gold standard in a sample of 40 children (median age: 4.3 years). The authors applied paired t-tests and correlation/regression analysis for comparing the smartphone values to the ECG values. Scatter plots were used to illustrate the correlations. R^2 values ranged between 0.071 and 0.857. In general, earlobe sensors outperformed sensors placed at toe or finger. The authors conclude that accuracy of the four apps is not sufficient for medical use.

Gao et al. (2016) introduced a flexible integrated sensor array (FISA) for a complex sweat analysis regarding lactate, glucose, sodium, potassium, and skin temperature. The authors analysed the chrono-electric and temperature-dependent behaviour of each sensor. The linear behaviour of each sensor (sensitivity, accuracy) was checked (relation of substrate concentration and current for glucose and lactate or relation of substrate concentration and potential for sodium and potassium) and confirmed experimentally. Precision was assessed by relative SD (metabolic sensors: ~1% SD; electrolyte sensors: ~5% SD). In addition, real-time monitoring was applied using specific ramped, continuous and graded load protocols on the cycle ergometer to test criterion validity. This study is a rare example where dynamic errors were addressed. The error curves show short latencies and overshoots depending on substance concentration.

Many sensors mentioned in Table 3 can either be deployed as stand-alone devices or they can be integrated in smartphones or wrist-worn devices such as smartwatches. Stand-alone sensors transmit data via wireless connectivity such as Bluetooth, WiFi, ANT or ZigBee (Mukhopadhyay, 2015; Kassal, Steinberg, & Steinberg, 2018).

Usually, the raw signals are further processed in order to quantify and qualify the relevant aspects of PA and FT. In addition, the results of the signal processing must be presented as visual, acoustic or haptic feedback to the user, e.g. regarding current state and discrepancies from target (Tang & Kay, 2017).

Data or signal processing requires more or less computational power depending on the complexity of algorithms and data volume. The different devices offer specific options for processing and presentation. On the one hand, small wrist-worn devices are light and non-obtrusive, but have low capacities for computation and presentation. On the other hand, smartphones offer enhanced capacities at the cost of higher weight and reduced wearing comfort. Due to the fact that small wrist-worn devices are preferred by users (e.g., Kettunen, Critchley & Kari, 2019), algorithms for low-capacity computations are often applied such as support-vector machines (SVM) or simple neural network implementations or other machine learning approaches (e.g., Zhou et al., 2018).

Technologies for PA and FT have to fulfil numerous quality criteria (see Table 5): At the sensor level, as has been mentioned above, the device has to be precise and accurate. Errors can result from the signal (e.g., signal-to-noise ratio), the sensor itself, sensor-movement interactions or algorithmic problems, i.e., error propagation. Generally, measurements and assessments should be objective, reliable and valid. In addition, often-used classifications like "correct" – "wrong" or "low" – "high" must be checked for sensitivity and selectivity (specificity). Unfortunately, no generally agreed standards for the assessment of measurement quality regarding PA or FT exist (e.g., Dowd et al., 2018).

Furthermore, to fulfil the functions specified in the previous sections, the system should have the following features (see also Preuschl et al. 2010; Fritz et al., 2014; Wiemeyer et al., 2016; Tang & Kay, 2017; Kari et al., 2016; Wiemeyer, 2018; Kettunen, Critchley & Kari, 2019):

- *Wireless communication* with devices to transmit, store and further analyse sensor data, e.g., with a database server.
- *Export and import function* for training schedule to and from external calendar systems
- *Import functions for media* like audios, texts, photos or videos
- Options for *synchronous and asynchronous communication*
- Options *for individualized visual, acoustic, and/or haptic (online or offline) feedback* regarding current status, discrepancies, training history (e.g., long-term trends and patterns), goal tracking (including flexible filtering, e.g., for relative versus absolute goal attainment), encouragement, appreciation, and reinforcement
- Options for *self-reflection and self-awareness*, e.g., analysis of training context and factors affecting training adherence, trends and patterns
- Connection to *social networks and communities*

Beyond the functionality of technology, mobile technologies for training should satisfy a number of further criteria, e.g., usability (effectiveness, efficiency, and satisfaction; ISO norm 9241-11:2016; Bevan et al., 2016, p.269). "Nowadays, usability is considered one of the most important aspects for the success of any technological product" (Paz & Pow-Sang, 2016, p.165). The top-5 methods for usability assessment comprising more than 70% of the applications are: questionnaires, user testing, heuristic evaluation, interview, and thinking aloud (Paz & Pow-Sang, 2016). The concept of user experiences denotes a broader approach than usability, including the (meta-)cognitive and emotional experiences of users when interacting with ICT within a specific context (Lallemand, Gronier & Koenig, 2015). Furthermore, for interactive ICT, certain "dialog principles" are defined by ISO norm 9241-110:2006, i.e. suitability for the task, self-descriptiveness, conformity with user expectations, suitability for learning, controllability, error tolerance, and suitability for individualization (Mentler & Herczeg, 2013, p.503).

Considering the fact, that many mobile fitness apps transfer data via wireless connectivity, security issues are also relevant to these applications. In this regard, Fereidooni et al. (2017) recently revealed severe security issues in the market-leading fitness tracker.

Beyond the "top-down" norms mentioned above, users themselves have specified numerous criteria for a positive experience of using mobile PA and FT apps. The most prominent expectations of users are (Casey et al., 2014; Kari et al., 2016; Tang et al., 2016; Wang et al., 2016; Kettunen, Critchley & Kari, 2019):

- ease of implementation and use
- clear, relevant, individualized and non-schematic information (presentation) matching the user's expectancies and capabilities
- ubiquitous and flexible support for effective and efficient training, awareness, self-reflection etc.
- fun and entertainment (particularly for teenagers; Kettunen & Kari, 2018)

To conclude this section, the following criteria for the evaluation of technology should be applied:

- *Sensors*: precision and accuracy, conditions of application (temperature, placement, movements etc.), errors (static – dynamic; random – systematic)
- *Data transmission and storage*: transmission rate, data security
- *Data processing*: algorithm (type, performance, feasibility)
- *Data classification*: validity, specificity and selectivity
- *Device*: technical features regarding computational capacity, presentation, connectivity, communication etc.
- *Device-user interaction*: usability (effectivity, efficiency, satisfaction; pragmatic and hedonic quality), dialogue and interaction (ISO norm 9241-110:2006)
- *Feasibility "in the wild"*, i.e., under prototypical PA and FT conditions: laboratory and real-life situations; systematic variation of relevant FITT parameters

Effects of mobile apps on PA and FT – evaluation and evidence

The criteria specified in the previous modules (training, BCT, and technology) can be considered as sine qua non regarding the final module: the actual impact of mobile apps on SPA and FT outcomes. Finally, the apps must on the one hand achieve the goals and outcomes they have been developed for and on the other hand find acceptance in the target group.

The gold standard for the summative evaluation of treatment effects is a two-arm randomized controlled trial (RCT) applying at least one pre- and post test with at least two parallel groups (e.g., Hecksteden et al., 2018). To allow for unbiased causal interpretations, group assignment should be randomized and method of randomization specified. Furthermore, sample composition must be explicitly planned and groups must be comparable regarding the most important features (e.g., gender, age, performance level, experience level; initial values of dependent variables). In addition, sample size should be carefully (pre-)determined to balance type 1 and 2 errors. At least, the personnel and testing persons should be blinded to the group assignment. The study should be based on clear hypotheses and outcomes must be measured or assessed according to the quality standards (reliability, validity, errors; precision, accuracy). Statistical procedures

must be appropriate regarding hypothesis, data scale as well as drop-outs and missing values. Despite the controversy regarding the use of inference statistics (e.g., Küberger et al., 2015), there is still good reason to apply (multivariate) Analysis of Variance ((M)ANOVA) to the data. Effect sizes should be reported as well as a priori and a posteriori estimation of statistical power. In longitudinal studies with fitness apps, dropouts and withdrawals as well as technical issues may cause missing data issues. However, to amend or at least mitigate this problem, appropriate statistical procedures have been proposed, e.g., multiple imputations or full information maximum likelihood (e.g., Lang & Little, 2018) or intention-to-treat analysis (e.g., McCoy, 2017).

Another option for a research design is a matched case-control trial (MCCT; Rose & Laan, 2009), where groups are matched regarding selected indicators. MCCTs allow for conditional rather causal interpretations and can mitigate confounding, but do not allow for avoiding bias. Furthermore, MCCT requires specific statistical procedures for analysis.

For exploratory purposes, feasibility or pilot studies can also be applied in order to prepare an RCT (Arain et al., 2010). However, these types of studies have numerous shortcomings, including lack of causal or conditional interpretation as well as lack of control over bias.

Due to the complexity of PA and FT (see also Table 3), training goals, the variety of possible training settings and the multi-dimensionality of human behaviour, a great variety of outcomes can be assessed (see Table 6), ranging from direct assessment of PA and fitness (e.g., Reilly et al., 2008; recent review: Dowd et al., 2018) over PA questionnaires (review: Helmerhorst et al., 2012) and self-reports to assessment of factors influencing outcomes such as attitude, motivation, training setting and quality of the app (e.g., Plonczynsky, 2000; McKay, Slykerman & Dunn, 2019).

In order to illustrate the procedure of summative evaluation, one selected best-practice example is described. The study of King et al. (2016) includes an 8-week RCT comparing three different mobile PA apps. The sample is determined according to explicit criteria (age, PA) and described in detail. Adequate recruitment procedures and random and blinded assignment are applied. Optimal sample size has been explicitly calculated. Furthermore, the apps have been explicitly developed and selected with regard to a theoretic background (motivation, social-cognitive theory, self-regulation). The intervention is controlled and described in detail, e.g., customized feedback, push and pull information components, baseline and application period. In addition, the research methods comprise objective primary outcomes (MVPA and sedentary time, derived from accelerometry) and secondary subjective outcomes (self-estimations, self-reports). However, further secondary outcomes such as anthropometry, physiological and psychic variables (see Table 6) have not been assessed. Statistical procedures (i.e., mixed-models analyses) are deliberately and adequately chosen. According to Romeo et al. (2019, additional material), risk of bias in this study is comparatively low. Finally, the report is comprehensive regarding participation (including dropouts), measurement (insufficient data) and descriptive statistics. However, regarding statistics, only significant results are reported in sufficient detail.

Table 6. Outcomes and operationalizations for the evaluation of fitness apps

Outcome/ Variable	Assessment (examples)
Steps per day, activity counts, METs, PAL, PA or MVPA duration/profile	On-body sensors (pedometer, accelerometer, heart rate monitor, armband)
	Observation, activity journal or log
	PA questionnaire
	Interview or calculation: Energy expenditure
Fitness	Various field and laboratory tests for different components of fitness, e.g., Cooper test, treadmill or bicycle ergometer tests
	Self-estimation questionnaire
	Interview
Anthropometry: Body mass, BMI, body composition	Scale
	Ruler, stadiometer
	Bioelectrical impedance analysis
Physiological: resting heart rate, blood pressure	Heart rate monitor
	Blood pressure devices
App usage: duration, usability	Recording
	Report (questionnaire, interview)
	Log, journal or diary
Psychic: goals, expectations, motivation, volition, emotion, self-efficacy, barriers, quality of life	Questionnaire
	Self-report
	Interview

For synthesis of research, systematic qualitative and quantitative methods are available. These methods should also follow the respective standards, e.g., PRISMA (Liberati et al., 2009), AMSTAR (Shea et al., 2009) or CONSORT standards (Guyatt et al., 2011). For example, the search procedure should be carefully documented. Furthermore, the quality of the primary studies (e.g., Maher et al., 2003) as well as risk of bias should be assessed (e.g., Higgins & Altman, 2008).

Regarding the impact of mobile apps on PA and FT, some reviews exist that will be addressed in chronological order to give a realistic impression of the possible impact of mobile fitness apps on SPA and PF.

Fanning, Mullen and McAuley (2012) performed a meta-analysis regarding the impact of early-state mobile applications (i.e., SMS/mobile phones and PDA) on PA. Search procedure and quality of the studies were documented. The 11 studies and 18 effect sizes yielded a weighted mean effect size of $g = 0.54$ (moderate). Moderator analysis revealed a significant large effect

regarding pedometer steps (as compared to MVPA duration) and a significant moderate effect of mobile phone (as compared to PDA).

Derbyshire and Dancey (2013) performed a systematic review regarding evidence for the impact of smartphones on females' health. Search procedure was documented according to the PRISMA standards. They identified only one study confirming the superiority of app-based self-monitoring over diary and website use.

O'Reilly and Sprujit-Metz (2013) performed a systematic review of mobile app interventions regarding PA. Literature search was carefully documented including data bases, search items, and selection procedure. In addition, the quality of the studies was rated according to an established instrument (Effective Public Health Practice Project Quality Assessment Tool). Nine of 12 studies reported significant positive effects on PA. Results regarding personalized information (tailoring) were inconsistent.

Stephens and Allen (2013) report a systematic review regarding the impact of mobile phone interventions on PA and weight loss. Literature search was carefully documented and seven studies were identified. Five studies applied text messaging and two studies used a smartphone app. Quality of studies was thoroughly discussed, four studies were classified as RCT. Five of seven studies reported at least one significant outcome in favour of mobile apps.

Bert et al. (2014) performed a literature search regarding the application of smartphones to health (particularly, nutrition, lifestyle and PA). Search was carefully documented and only one matched case-control trial was located. Three studies were just RCT study protocols (without results), the rest included application descriptions and validations.

Mateo et al. (2015) performed a review and meta-analysis of studies applying mobile phone apps for weight loss and increase of PA. Search was carefully documented according to the PRISMA standard and seven studies addressing PA were identified (6 RCT, 1 MCCT). Mean standardized effect size was low to moderate (0.40) and not significant. Risk of bias was assessed based on the Cochrane standard. For the seven studies, one to four items were considered critical. All studies suffered from performance bias, while 6 studies suffered from detection bias.

Matthews et al. (2016) reviewed 20 studies promoting PA. Search was carefully documented. Studies included 11 RCT or outcome studies, 5 studies focussed on software design and evaluation and the remaining 4 studies targeted stakeholders' opinions using focus groups and interviews. According to the Persuasive Systems Design (PSD) model (Oinas-Kukkonen & Harjumaa, 2009), the authors identified four types of support: primary task support (e.g., self-monitoring, tailoring, or personalization), dialogue support (e.g., suggestions, reminders, praises, and rewards), social support (e.g., social comparison, social learning, and competition), and system credibility support (e.g., authority and trustworthyness).

McKay et al. (2018) identified two effectiveness studies for PA. The only RCT has already been covered by the review of Mateo et al. (2015).

Romeo et al. (2019) performed a meta-analysis regarding the impact of smartphone apps on PA (steps per day). Literature search as well as quality of the studies and app features were carefully documented. The authors found a non-significant effect size for the 6 studies ($SMD = 0.21$).

Therefore, existing reviews are inconclusive, indicating no or low to moderate effects of fitness apps. Furthermore, only few high-quality studies exist.

To conclude this section, evaluation of mobile apps targeting PA and FT should adopt the gold standard of RCT meeting established quality criteria as well as avoiding risk of bias. In sum, the following aspects have to be adequately considered for single studies:

- Research design: RCT or MCCT
- Sample: adequate structure, size, and characteristics
- Treatment: according to theory, model, and hypotheses; clear plan; randomized assignment
- Dependent variables: according to intended effects (level); blinding of assessors; validity criteria
- Statistical procedures: according to design and data quality; intention-to-treat, multiple imputation
- Risk of bias – five core areas (Higgins & Altman, 2008): selection, performance, attrition, detection, and reporting.
- Research report: according to the established standards

Discussion

Currently, many mobile applications are available that offer support for SPA and FT. However, the quality of these applications has not yet been addressed sufficiently. Rather, with only few exceptions, either low-quality research designs have been applied or only selective aspects of the applications have been tested like application of BCT, technical quality or impact on SPA and FT. What is still missing, is an approach that integrates the most important aspects and disciplinary insights contributing to the outcome quality of mobile applications for healthy people.

Therefore, the main contribution of this paper is the introduction of a modular interdisciplinary framework for the evaluation of mobile applications aimed to support SPA and FT in health people with low training level. The framework comprises four modules: Training, BCT, technology and sensors, and evaluation of effects (see Figure 3).

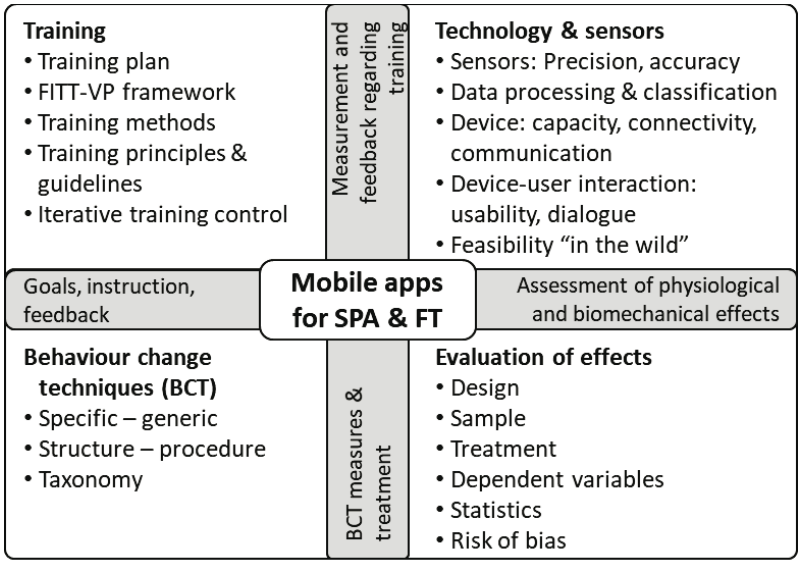


Figure 3. TBTE framework – Overview; legend: FITT-VP – frequency, intensity, time, type, volume, progression; SPA – sustainable physical activity; FT – fitness training

The four modules and the respective specifications can provide a comprehensive guidance for the systematic (formative or summative) evaluation of mobile applications dedicated to SPA and FT (see Table 7). The specifications have been derived from insights in the relevant scientific fields, for example, training science, exercise physiology, social psychology, computer science, and civil engineering. Therefore, the appropriate scientific standards have been considered. Of course, the list of criteria can be extended or shortened according to the specific goal of the evaluation. However, the proposed framework is considered a reasonable starting point for a comprehensive evaluation of mobile fitness apps in the specified target group.

Table 7. Overview of criteria in the four modules

Training	**Behavior change techniques**	**Technology & sensors**	**Evaluation**
Individual assessment	Action planning	Sensors: reliability & errors	Research design
Individual and flexible training plans	Feedback, reinforcement and rewards	Data transmission and storage: rate, security	Sample
Data base: fitness exercises	Informational guidance	Data classification: validity, specificity and selectivity	Treatment & procedure
Data base: tests and diagnostics	Prompts, cues, and reminders	Device: technical features	Dependent variables
Recording and feedback	Social comparisons and competition	Device-user interaction: usability, dialogue and interaction	Statistical procedures
Analysis and presentation of training data	Social support	Feasibility “in the wild”	Risk of bias
Online and offline feedback	Supporting intrinsic motivation, self-efficacy and self-concordance	Export and import functions (data bases, media)	Standard research report
Instructions for individualized training		Communication functions	
Coaching functions		Customization	

Note that there is considerable overlap of the four modules (see Figure 3). For example, technology and sensors have to take into account the specific aspects of training, e.g., relevant methods and parameters that have to be assessed. In a similar vein, BCT and evaluation are closely related, since evaluation has to take into account the specific BCT and training concepts, e.g., when selecting and operationalizing adequate dependent measures.

Regarding the concrete application of the framework, numerous options are available, for example, regarding the selection of sensors and research methods. This "freedom of choice" has been addressed in all sections, e.g., regarding the selection of training methods in the first section, the selection of BCT approaches in the second section, the selection of dimensions and parameters for assessment in the third section, and the selection of evaluation methods in the fourth section. However, choice has consequences regarding quality of outcome, application process, technology, and evaluation.

Finally, the proposed model is selective and does not cover all the insights from the relevant scientific disciplines. Rather, the framework is meant to be open to adding further specification that may be deemed important under certain conditions.

Conclusion

The evaluation of the quality of mobile applications supporting SPA and FT requires an interdisciplinary approach that integrates the insights from the relevant scientific disciplines. In this paper, we propose an interdisciplinary framework that integrates the insights from trainings science, exercise physiology, social psychology, computer science, and civil engineering. The framework consists of four modules: training, BCT, technology and sensors and evaluation of effects.

Considering this framework may contribute to enhance the quality of evaluating mobile applications aiming at SPA and FT in healthy person with low engagement in PA and low performance level.

Acknowledgement

The author would like to thank the reviewers for their careful, responsible and constructive feedback as well as their endurance and patience that resulted in many helpful comments and recommendations to substantially increase the quality of the paper.

References

Abraham, C. & Michie, S. (2008). A taxonomy of behavior change techniques used in interventions. *Health Psychology, 27* (3), 379–387.

ACSM (2011). Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: Guidance for prescribing exercise. *Medicine & Science in Sports & Exercise, 43* (7), 1334-1359.

Ainsworth, B. E., Haskell, W. L., Leon, A. S., Jacobs, J. D., Montoye, H. J., Sallis, J. F., & Paffenbarger, J. R. (1993). Compendium of physical activities: Classification of energy costs of human physical activities. *Medicine and science in sports and exercise*, *25* (1), 71-80.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50* (2), 179-211.

Arain, M., Campbell, M. J., Cooper, C. L., & Lancaster, G. A. (2010). What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC medical research methodology*, *10* (1), 67.

Arem, H., Moore, S. C., Patel, A., Hartge, P., De Gonzalez, A. B., Visvanathan, K., ... & Linet, M. S. (2015). Leisure time physical activity and mortality: a detailed pooled analysis of the dose-response relationship. *JAMA internal medicine*, *175* (6), 959-967.

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports medicine*, *26* (4), 217-238.

Baca, A. (2015). Data acquisition and processing. In A. Baca (ed.), *Computer Science in Sport: Research and practice* (pp.46-81). London: Routledge.

Bandura, A. (1999). A social cognitive theory of personality. In L. Pervin & O. John (Ed.), *Handbook of personality* (2nd ed., pp. 154-196). New York: Guilford Publications.

Batacan, R. B., Duncan, M. J., Dalbo, V. J., Tucker, P. S., & Fenning, A. S. (2017). Effects of high-intensity interval training on cardiometabolic health: a systematic review and meta-analysis of intervention studies. *British Journal of Sports Medicine*, *51*(6), 494-503.

Battenberg, A. K., Donohoe, S., Robertson, N., & Schmalzried, T. P. (2017). The accuracy of personal activity monitoring devices. *Seminars in Arthroplasty, 28* (2), 71-75.

Bender, C. G., Hoffstot, J. C., Combs, B. T., Hooshangi, S., & Cappos, J. (2017). Measuring the fitness of fitness trackers. In *Sensors Applications Symposium (SAS), 2017 IEEE* (pp. 1-6). New York, NY: IEEE.

Bert, F., Giacometti, M., Gualano, M. R., & Siliquini, R. (2014). Smartphones and health promotion: A review of the evidence. *Journal of medical systems, 38* (1), 1-11.

Bevan, N., Carter, J., Earthy, J., Geis, T., & Harker, S. (2016). New ISO standards for usability, usability reports and usability measures. In *International Conference on Human-Computer Interaction* (pp. 268-278). Cham: Springer.

Bondaronek, P., Alkhaldi, G., Slee, A., Hamilton, F. L., & Murray, E. (2018). Quality of publicly available physical activity apps: Review and content analysis. *JMIR mHealth and uHealth*, *6*(3), e53.

Borg, G. (1998). *Borg´s perceived exertion and pain scales*. Champaign, II.: Human Kinetics.

Case, M. A., Burwick, H. A., Volpp, K. G., & Patel, M. S. (2015). Accuracy of smartphone applications and wearable devices for tracking physical activity data. *Journal of the American Medical Association*, *313* (6), 625-626.

Casey, M., Hayes, P. S., Glynn, F., ÓLaighin, G., Heaney, D., Murphy, A. W., & Glynn, L. G. (2014). Patients' experiences of using a smartphone application to increase physical activity: The SMART MOVE qualitative study in primary care. *British Journal of General Practice, 64* (625), e500-e508.

Caspersen, C. J., Powell, K. E., & Christenson, G. M. (1985). Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public health reports*, *100* (2), 126.

Champion, V. L. & Skinner, C. S. (2008). The health belief model. In K. Glanz, B.K. Rimer & K. Viswanath (eds.), *Health behavior and health education: Theory, research, and practice* (pp. 45-65). San Francisco, CA: Wiley.

Chi-Wai, R. K., Sai-Chuen, S. H., So-Ning, T. M., Ka-Shun, P. W., Wing-Kuen, K. L., & Choi-Ki, C. W. (2011). Can mobile virtual fitness apps replace human fitness trainer? In *The*

5th International Conference on New Trends in Information Science and Service Science (Vol. 1, pp. 56-63). New York, NY: IEEE.

Conroy, D. E., Yang, C. H., & Maher, J. P. (2014). Behavior change techniques in top-ranked mobile apps for physical activity. *American journal of preventive medicine*, *46* (6), 649-652.

Derbyshire, E. & Dancey, D. (2013). Smartphone medical applications for women's health: What is the evidence-base and feedback? *International journal of telemedicine and applications*, Article ID 782074.

Direito, A., Dale, L. P., Shields, E., Dobson, R., Whittaker, R., & Maddison, R. (2014). Do physical activity and dietary smartphone applications incorporate evidence-based behaviour change techniques? *BMC Public Health, 14* (646), 1-7.

Donabedian, A. (1988). The quality of care: How can it be assessed? *JAMA, 260*, 1743-1748.

Dowd, K. P., Szeklicki, R., Minetto, M. A., Murphy, M. H., Polito, A., Ghigo, E., ... & Tomczak, M. (2018). A systematic literature review of reviews on techniques for physical activity measurement in adults: a DEDIPAC study. *International Journal of Behavioral Nutrition and Physical Activity*, *15* (1), 15.

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.

Fanning, J., Mullen, S. P., & McAuley, E. (2012). Increasing physical activity with mobile devices: A meta-analysis. *Journal of medical Internet research*, *14*(6).

Farrow, D., & Robertson, S. (2017). Development of a skill acquisition periodisation framework for high-performance sport. *Sports Medicine, 47* (6), 1043–1054.

Fereidooni, H., Classen, J., Spink, T., Patras, P., Miettinen, M., Sadeghi, A. R., Hollick, M., & Conti, M. (2017). Breaking fitness records without moving: Reverse engineering and spoofing fitbit. In *International Symposium on Research in Attacks, Intrusions, and Defenses* (pp. 48-69). Cham: Springer.

Fokkema, T., Kooiman, T. J., Krijnen, W. P., Schans, C. P. van der, & Groot, M. de (2017). Reliability and validity of ten consumer activity trackers depend on walking speed. *Medicine and science in sports and exercise*, *49* (4), 793-800.

Fritz, T., Huang, E. M., Murphy, G. C., & Zimmermann, T. (2014,). Persuasive technology in the real world: A study of long-term use of activity sensing devices for fitness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 487-496). New York, NY: ACM.

Fröhlich, M., Müller, F., Schmidtbleicher, D. & Emrich, E. (2009). Outcome-Effekte verschiedener Periodisierungsmodelle im Krafttraining. *Deutsche Zeitschrift für Sportmedizin, 60* (10), 307-314.

Fuchs, R., Goehner, W., & Seelig, H. (2011). Long-term effects of a psychological group intervention on physical exercise and health: The MoVo concept. *Journal of Physical Activity and Health*, *8* (6), 794-803.

Fuchs, R., Seelig, H., Göhner, W., Burton, N. W., & Brown, W. J. (2012). Cognitive mediation of intervention effects on physical exercise: Causal models for the adoption and maintenance stage. *Psychology & health*, *27* (12), 1480-1499.

Gao, W., Emaminejad, S., Nyein, H. Y. Y., Challa, S., Chen, K., Peck, A., Fahad, H.M., Ota, H., Shiraki, H., Kiriya, D., Lien, D.-H., Brooks, G.A., Davis, R.W., & Javey, A. (2016). Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. *Nature*, *529* (7587), 509.

Gibson, A. L., Wagner, D., & Heyward, V. (2018). *Advanced Fitness Assessment and Exercise Prescription* (8th edition). Champaign, Ill.: Human kinetics.

Guissard, N., Duchateau, J. & Hainaut, K. (1988). Muscle stretching and motoneuron excitability. *European Journal of Applied Physiology, 58*, 47-52.

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Susan Norris, S., Falck-Ytter, Y., Glasziou, P., deBeer, H., Jaeschke, R., Rind, D., Meerpohl, J., Dahm, P., & Schünemann, H. J. (2011). GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology*, *64* (4), 383-394.

Hagger, M. S. & Chatzisarantis, N. L. (2014). An integrated behavior change model for physical activity. *Exercise and Sport Sciences Reviews*, *42* (2), 62-69.

Halson, S. L., Peake, J. M., & Sullivan, J. P. (2016). Wearable technology for athletes: Information overload and pseudoscience? *International Journal of Sports Physiology and Performance, 11*, 705 -706.

He, Y., & Li, Y. (2013). Physical Activity Recognition Utilizing the Built-In Kinematic Sensors of a Smartphone. *International Journal of Distributed Sensor Networks*, *2013,* Article ID 481580.

Heckhausen, H. (1989). *Motivation und Handeln* (2nd ed.). [Motivation and action] Berlin: Springer.

Hecksteden, A., Faude, O., Meyer, T., & Donath, L. (2018). How to construct, conduct and analyze an exercise training study? *Frontiers in physiology, 9*, 1007.

Heikenfeld, J., Jajack, A., Rogers, J., Gutruf, P., Tian, L., Pan, T., Li, R., Khine, M., Kim, J., Wang, J., & Kim, J. (2018). Wearable sensors: Modalities, challenges, and prospects. *Lab on a Chip*, *18* (2), 217-248.

Helmerhorst, H. H. J., Brage, S., Warren, J., Besson, H., & Ekelund, U. (2012). A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. *International Journal of Behavioral Nutrition and Physical Activity*, *9* (1), 103.

Higgins, J. P. & Altman, D. G. (2008). Assessing risk of bias in included studies. In J.P. Higgins & S. Green (eds.), *Cochrane handbook for systematic reviews of interventions: Cochrane book series* (pp. 187-241). Chichester: Wiley-Blackwell.

Higgins, J.P. & Green, S. (eds.). (2008). *Cochrane handbook for systematic reviews of interventions: Cochrane book series*. Chichester: Wiley-Blackwell.

Ho, C. L., Fu, Y. C., Lin, M. C., Chan, S. C., Hwang, B., & Jan, S. L. (2014). Smartphone applications (apps) for heart rate measurement in children: Comparison with electrocardiography monitor. *Pediatric cardiology*, *35* (4), 726-731.

Hohmann, A., Lames, M. & Letzelter, M. (2002). *Einführung in die Trainingswissenschaft.* [Introduction to training science] Wiebelsheim: Limpert.

Janssen, I., & LeBlanc, A. G. (2010). Systematic review of the health benefits of physical activity and fitness in school-aged children and youth. *International journal of behavioral nutrition and physical activity*, *7* (1), 40.

Kari, T., Koivunen, S., Frank, L., Makkonen, M., & Moilanen, P. (2016). Critical experiences during the implementation of a self-tracking technology. In *PACIS 2016: Proceedings of the 20th Pacific Asia Conference on Information Systems* (pp. 129-144). Association for Information Systems. Retrieved from http://aisel.aisnet.org/pacis2016/129/

Kari, T. & Rinne, P. (2018). Influence of digital coaching on physical activity: Motivation and behaviour of physically inactive individuals. In A. Pucihar, M. Kljajič, P. Ravesteijn, J. Seitz, & R. Bons (Eds.), *Bled 2018: Proceedings of the 31th Bled eConference. Digital Transformation: Meeting the Challenges* (pp. 127-145). Maribor: University of Maribor Press.

Kassal, P., Steinberg, M. D., & Steinberg, I. M. (2018). Wireless chemical sensors and biosensors: A review. *Sensors and Actuators B: Chemical*, *266*, 228.

Kellmann, M. & Kallus, K. W. (2001). *Recovery-stress questionnaire for athletes: User manual* (Vol. 1). Champaign, Il.: Human Kinetics.

Kellmann, M., Bertollo, M., Bosquet, L., Brink, M., Coutts, A. J., Duffield, R., ... & Kallus, K. W. (2018). Recovery and performance in sport: consensus statement. *International journal of sports physiology and performance*, *13* (2), 240-245.

Kendzierski, D. & DeCarlo, K. J. (1991). Physical activity enjoyment scale: Two validation studies. *Journal of sport and exercise psychology*, *13* (1), 50-64.

Kettunen, E., Critchley, W., & Kari, T. (2019). Can digital coaching boost your performance? A qualitative study among physically active people. In *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS 2019)* (pp. 1331-1340). University of Hawai'i at Manoa. Retrieved April 4, 2019 from http://hdl.handle.net/10125/59574

Kettunen, E. & Kari, T. (2018). Can sport and wellness technology be my personal trainer? Teenagers and digital coaching. In A. Pucihar, M. Kljajič, P. Ravesteijn, J. Seitz, & R. Bons (Eds.), *Bled 2018: Proceedings of the 31th Bled eConference. Digital Transformation: Meeting the Challenges* (pp. 463-476). Maribor: University of Maribor Press.

Khaylis, A., Yiaslas, T., Bergstrom, J., & Gore-Felton, C. (2010). A review of efficacious technology-based weight-loss interventions: five key components. *Telemedicine and e-Health*, *16* (9), 931-938.

King, A. C., Hekler, E. B., Grieco, L. A., Winter, S. J., Sheats, J. L., Buman, M. P., ... & Cirimele, J. (2016). Effects of three motivationally targeted mobile device applications on initial physical activity and sedentary behavior change in midlife and older adults: a randomized trial. *PloS one*, *11*(6), e0156370.

Knight, E., Stuckey, M. I., Prapavessis, H., & Petrella, R. J. (2015). Public health guidelines for physical activity: Is there an app for that? A review of android and apple app stores. *JMIR mHealth and uHealth*, *3* (2).

Kooiman, T. J., Dontje, M. L., Sprenger, S. R., Krijnen, W. P., van der Schans, C. P., & de Groot, M. (2015). Reliability and validity of ten consumer activity trackers. *BMC sports science, medicine and rehabilitation*, *7* (1), 24.

Kranz, M., Möller, A., Hammerla, N., Diewald, S., Plötz, T., Olivier, P., & Roalter, L. (2013). The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices. *Pervasive and Mobile Computing*, *9* (2), 203-215.

Kühberger, A., Fritz, A., Lermer, E., & Scherndl, T. (2015). The significance fallacy in inferential statistics. *BMC research notes*, *8* (1), 84.

Lachman, M. E., Lipsitz, L., Lubben, J., Castaneda-Sceppa, C., & Jette, A. M. (2018). When adults don't exercise: Behavioral strategies to increase physical activity in sedentary middle-aged and older adults. *Innovation in aging*, *2*(1), igy007.

Lallemand, C., Gronier, G., & Koenig, V. (2015). User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in Human Behavior*, *43*, 35-48.

Lang, K. M., & Little, T. D. (2018). Principled missing data treatments. *Prevention Science*, *19*(3), 284-294.

Leunes, A. & Burger, J. (2000). Profile of mood states research in sport and exercise psychology: Past, present, and future. *Journal of applied sport psychology*, *12* (1), 5-15.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS medicine*, *6* (7), e1000100.

Ludwig, M., Hoffmann, K., Endler, S., Asteroth, A., & Wiemeyer, J. (2018). Measurement, prediction, and control of individual heart rate responses to exercise—Basics and options for wearable devices. *Frontiers in physiology*, *9*, 778.

Maher, C. G., Sherrington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical therapy*, *83*(8), 713-721.

Marshall, S. J. & Biddle, S. J. (2001). The transtheoretical model of behavior change: A meta-analysis of applications to physical activity and exercise. *Annals of behavioral medicine*, *23* (4), 229-246.

Mateo, G. F., Granado-Font, E., Ferré-Grau, C., & Montaña-Carreras, X. (2015). Mobile phone apps to promote weight loss and increase physical activity: A systematic review and meta-analysis. *Journal of medical Internet research*, *17* (11), e253.

Matthews, J., Win, K. T., Oinas-Kukkonen, H., & Freeman, M. (2016). Persuasive technology in mobile applications promoting physical activity: A systematic review. *Journal of medical systems*, *40* (3), 72.

McCoy, C. E. (2017). Understanding the intention-to-treat principle in randomized controlled trials. *Western Journal of Emergency Medicine*, *18* (6), 1075.

McKay, F. H., Cheng, C., Wright, A., Shill, J., Stephens, H., & Uccellini, M. (2018). Evaluating mobile phone applications for health behaviour change: A systematic review. *Journal of telemedicine and telecare*, *24* (1), 22-30.

McKay, F. H., Slykerman, S., & Dunn, M. (2019). The App Behavior Change Scale: Creation of a scale to assess the potential of apps to promote behavior change. *JMIR mHealth and uHealth*, *7* (1), e11130.

Mentler, T. & Herczeg, M. (2013). Applying ISO 9241-110 dialogue principles to tablet applications in emergency medical services. In *Proceedings of the 10th International ISCRAM Conference – Baden-Baden, Germany, May 2013* (pp.502-506). Baden-Baden: ISCRAM (http://www.iscram.org/content/iscram2013-academic-papers)

Michie, S., Ashford, S., Sniehotta, F. F., Dombrowski, S. U., Bishop, A., & French, D. P. (2011). A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: The CALO-RE taxonomy. *Psychology & Health*, *26* (11), 1479-1498.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M. P., Cane, J., & Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine*, *46* (1), 81-95.

Mukhopadhyay, S. C. (2015). Wearable sensors for human activity monitoring: A review. *IEEE sensors journal*, *15* (3), 1321-1330.

Munson, S.A. & Consolvo, S. (2012). Exploring goal-setting, rewards, self-monitoring, and sharing to motivate physical activity. In *2012 6th international conference on pervasive computing technologies for healthcare (pervasive health) and workshops* (pp. 25-32). New York, NY: IEEE.

Oinas-Kukkonen, H. & Harjumaa, M. (2009). Persuasive systems design: Key issues, process model and system features. *Communications of the Association for Information Systems, 24* (1)*,* 28.

O'Donovan, G., Blazevich, A. J., Boreham, C., Cooper, A. R., Crank, H., Ekelund, U., ... & Hamer, M. (2010). The ABC of Physical Activity for Health: a consensus statement from the British Association of Sport and Exercise Sciences. *Journal of sports sciences*, *28*(6), 573-591.

O'Reilly, G. A. & Spruijt-Metz, D. (2013). Current mHealth technologies for physical activity assessment and promotion. *American journal of preventive medicine, 45* (4), 501-507.

Paz, F. & Pow-Sang, J. A. (2016). A systematic mapping review of usability evaluation methods for software development process. *International Journal of Software Engineering and Its Applications*, *10* (1), 165-178.

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.

Pelletier, L. G., Tuson, K. M., Fortier, M. S., Vallerand, R. J., Briere, N. M., & Blais, M. R. (1995). Toward a new measure of intrinsic motivation, extrinsic motivation, and amotivation in sports: The Sport Motivation Scale (SMS). *Journal of sport and Exercise Psychology*, *17* (1), 35-53.

Plonczynski, D. J. (2000). Measurement of motivation for exercise. *Health Education Research*, *15*(6), 695-705.

Poitras, V. J., Gray, C. E., Borghese, M. M., Carson, V., Chaput, J. P., Janssen, I., Katzmarzyk, P. T., Pate, R. R., Gorber, S. C., Kho, M. E., Sampson, M., & Tremblay, M.S. (2016). Systematic review of the relationships between objectively measured physical activity and health indicators in school-aged children and youth. *Applied Physiology, Nutrition, and Metabolism*, *41* (6), S197-S239.

Preuschl, E., Baca, A., Novatchkov, H., Kornfeind, P., Bichler, S., & Boecskoer, M. (2010). Mobile motion advisor – A feedback system for physical exercise in schools. *Procedia Engineering*, *2* (2), 2741-2747.

Prochaska, JO, Redding, CA, & Evers, K. (2008). The transtheoretical model and stages of change. In K. Glanz, F.M. Lewis, & B.K. Rimer (Eds.), *Health behavior and health education* (4th ed., pp.97-121). San Francisco: Jossey-Bass.

Reilly, J. J., Penpraze, V., Hislop, J., Davies, G., Grant, S., & Paton, J. Y. (2008). Objective measurement of physical activity and sedentary behaviour: review with new data. *Archives of disease in childhood*, *93* (7), 614-619.

Rhea, C. K., Felsberg, D. T., & Maher, J. P. (2018). Toward Evidence-Based Smartphone Apps to Enhance Human Health: Adoption of Behavior Change Techniques. *American Journal of Health Education*, *49*(4), 210-213.

Roda, A., Michelini, E., Zangheri, M., Di Fusco, M., Calabria, D., & Simoni, P. (2016). Smartphone-based biosensors: A critical review and perspectives. *TrAC Trends in Analytical Chemistry*, *79*, 317-325.

Romeo, A., Edney, S., Plotnikoff, R., Curtis, R., Ryan, J., Sanders, I., ... & Maher, C. (2019). Can Smartphone Apps Increase Physical Activity? Systematic Review and Meta-Analysis. *Journal of medical Internet research*, *21* (3), e12053.

Rose, S. & Laan, M. J. van der (2009). Why match? Investigating matched case-control study designs with causal effect estimation. *The international journal of biostatistics*, *5* (1), Article 1.

Ryan, R. M. & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55* (1), 68-78.

Schmidt, B., Benchea, S., Eichin, R., & Meurisch, C. (2015). Fitness tracker or digital personal coach: How to personalize training. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (pp. 1063-1067). New York, NY: ACM.

Shameli, A., Althoff, T., Saberi, A., & Leskovec, J. (2017). How gamification affects physical activity: Large-scale analysis of walking challenges in a mobile application. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 455-463). Geneva: International World Wide Web Conferences Steering Committee.

Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J., Henry, D.A., & Boers, M. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of clinical epidemiology*, *62* (10), 1013-1020.

Stephens, J., & Allen, J. (2013). Mobile phone interventions to increase physical activity and reduce weight: A systematic review. *The Journal of cardiovascular nursing*, *28* (4), 320.

Tang, L. M., Day, M., Engelen, L., Poronnik, P., Bauman, A., & Kay, J. (2016). Daily & hourly adherence: Towards understanding activity tracker accuracy. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3211-3218). New York, NY: ACM.

Tang, L. M. & Kay, J. (2017). Harnessing long term physical activity data – How long-term trackers use data and how an adherence-based interface supports new insights. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1* (2), Article 26.

Teixeira, P. J., Carraça, E. V., Markland, D., Silva, M. N., & Ryan, R. M. (2012). Exercise, physical activity, and self-determination theory: A systematic review. *International journal of behavioral nutrition and physical activity*, *9* (1), 78.

Toigo, M., & Boutellier, U. (2006). New fundamental resistance exercise determinants of molecular and cellular muscle adaptations. *European journal of applied physiology*, *97* (6), 643-663.

Wackerhage, H., Schoenfeld, B. J., Hamilton, D. L., Lehti, M., & Hulmi, J. J. (2018). Stimuli and sensors that initiate skeletal muscle hypertrophy following resistance exercise. *Journal of Applied Physiology*, *126* (1), 30-43.

Wagner, P. (2000). *Aussteigen oder Dabeibleiben?* [Get off or stay?] Darmstadt: WBG.

Wahl, Y., Düking, P., Droszez, A., Wahl, P., & Mester, J. (2017). Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions. *Frontiers in physiology*, *8*, 725.

Wang, J. B., Cataldo, J. K., Ayala, G. X., Natarajan, L., Cadmus-Bertram, L. A., White, M. M., Madanat, H., Nichols, J. F., & Pierce, J. P. (2016). Mobile and wearable device features that matter in promoting physical activity. *Journal of mobile technology in medicine*, *5* (2), 2-11.

Warraich, M. U. (2016). Wellness routines with wearable activity trackers: A systematic review. In *MCIS 2016 Proceedings* (Article 35). Paphos, Cyprus: http://aisel.aisnet.org/mcis2016/.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of personality and social psychology*, *54* (6), 1063-1070.

WHO (2010). *Global recommendations on physical activity for health*. Geneva: WHO.

WHO (2018). *More active people for a healthier world. Global action plan on physical activity 2018-2030*. Geneva: WHO.

Wiemeyer, J. (2018). Fitness Apps – Was erwarten die User? [Fitness apps – What are the users' expectations?] In D. Link, A. Hermann, M. Lames & V. Senner (eds.), *Sportinformatik XII* (pp. 90-91). Hamburg: Feldhaus-Czwalina.

Wiemeyer, J., Hatzky, W., Henrich, J. & Seelert, P. (2016). Modern – Mobil – Motivierend = Effektiver & Effizienter? Eine kritische Analyse ausgewählter mobiler Trainings-Applikationen. [Modern – mobile – motivating = more effective and more efficient? A critical analysis of selected applications for mobile training] In K. Witte & J. Edelmann-Nusser (eds.), *Sportinformatik XI.* (pp.29-34). Aachen: Shaker.

Williams, S. L. & French, D. P. (2011). What are the most effective intervention techniques for changing physical activity self-efficacy and physical activity behaviour – and are they the same? *Health Education Research*, *26* (2), 308-322.

Wong, C., Zhang, Z. Q., Lo, B., & Yang, G. Z. (2015). Wearable sensing for solid biomechanics: A review. *IEEE Sensors Journal*, *15* (5), 2747-2760.

Yang, C. H., Maher, J. P., & Conroy, D. E. (2015). Implementation of behavior change techniques in mobile applications for physical activity. *American journal of preventive medicine, 48* (4), 452-455.

Yang, R., Shin, E., Newman, M. W., & Ackerman, M. S. (2015). When fitness trackers don't 'fit': End-user difficulties in the assessment of personal tracking device accuracy. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 623-634). New York, NY: ACM.

Zhou, M., Fukuoka, Y., Mintz, Y., Goldberg, K., Kaminsky, P., Flowers, E., & Oi, A. (2018). Evaluating machine learning–based automated personalized daily step goals delivered through a mobile phone app: Randomized controlled trial. *JMIR mHealth and uHealth*, *6* (1), e28.