Wearable Activity Tracker Literature Review (January 2009 – July 2016)

Shawn E. Soutiere, Brennan D. Cox, Melissa D. Laird, Rachel R. Markwald, Jay H. Heaney, Evan D. Chinoy, & Rita G. Simmons

Naval Health Research Center, San Diego, California

Naval Health Research Center

Naval Health Research Center
140 Sylvester Road
San Diego, CA 92106-3521

REVIEWED AND APPROVED BY:

_____	03/27/2017
CAPT Rita G. Simmons, Ph.D., MSC, USN	Date
Commanding Officer	
Naval Health Research Center San Diego	
140 Sylvester Rd.	
San Diego, CA 92106-3521	

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Background/Objectives

Wearable activity trackers (WATs) have become increasingly popular for their purported ability to provide users with real-time, tailored information about their daily health-related activities. These devices are portrayed as user-friendly and beneficial to well-being, which raises the question of their potential application for health promotion among military populations. However, the rapid evolution of WAT technology is outpacing efforts to test, evaluate, and validate these devices. Therefore, the purpose of this systematic review was to summarize the evidence on WAT reliability and validity for measuring physical activity (PA) and sleep, and the utility of incorporating WATs in behavior modification programs.

Methods

Three distinct searches were conducted using PubMed, Ovid, Scopus, and Google Scholar. The searches identified 317 full-length studies published in English language journals from January 2009 to July 2016. After reviewing for inclusion and exclusion criteria, 34 studies were retained. Only studies of commercially available devices that reported objective measures were included. Studies unrelated to PA or that focused on specific clinical populations were excluded.

Results

Fifteen studies contained validity and/or reliability data. Overall, the results indicated that WAT reliability and validity varied considerably by task, measure, and device. Standard WAT measurements included step count, time spent in PA, energy expenditure, and heart rate. Step count was the most frequently reported metric (12 studies covering 26 devices) and had the highest correlations to actual steps taken for walking, jogging, or running (r = .67 to .97). All trackers, however, significantly underestimated actual steps when either the upper or lower body was rigid (carrying/pushing), stationary, and/or when the participant did not exhibit a "normal" gait pattern (mean absolute percent error [MAPE] = 3 to 33%). No studies (two studies covering four devices) accurately measured time spent in moderate to vigorous PA during free-living conditions (MAPE = 52 to 92%). Energy expenditure was underestimated by six of eight devices and overestimated by two of eight devices (MAPE = 11.2 to 83.4%). WAT-measured heart rate (five devices) was reviewed in two reports and correlations to reference measures were moderate (r = .67) to excellent (r = .95).

Fifteen studies contained sleep and activity data, including eight original research papers and seven literature reviews. Among the review articles, the consensus was WATs with sleep-tracking features showed theoretical promise but had not yet demonstrated sufficient validity. WAT validity was assessed by comparison with gold standards. Six of eight studies found WATs overestimated total sleep time (range: 8.0 to 67.1 minutes). Five of six studies found overestimations in sleep efficiency (range: 1.8 to 14.5%). Five studies evaluated wake after sleep onset (WASO) and performed sensitivity analyses for sleep; four of which found WATs underestimated WASO (range: 5.6 to 32.2 minutes). Five of five studies showed high sensitivity

for agreement of sleep between WATs and gold standards (range: 92 to 98%); however, specificity was low (range: 20 to 66%) indicating WATs misclassified wakefulness as sleep.

Four studies contained experimental data on WAT-based behavior change interventions. Methodologies differed, making the results of these studies incomparable. Overall, WATs with coaching/personal intervention showed short-term gains in PA, though long-term benefits have not yet been demonstrated. In one study, physically inactive participants issued WATs self-reported increased time spent in PA over baseline (74 minutes/week vs. 216 minutes/week); these gains were maintained for 6 months after WAT use discontinued. In three other studies, the effects of a WAT alone were compared with additional behavior change techniques. In one, only three of seven overweight subjects issued WATs increased activity levels (step count); however, after adding individualized coaching, all subjects realized significant gains (10.2 to 89.1% increase in steps). Another study found subjects who received a WAT plus individualized coaching lost more weight (mean change ± SE = –6.40 ± 1.17 kg; –7.37 ± 1.29%) than those who received a WAT alone (–4.04 ± 1.37 kg; –4.35 ± 1.29%). However, in the fourth study, overweight subjects reported WATs combined with automated, non-personalized fitness prompts showed no change in activity over a WAT alone. A recent study reported on overweight workers who received WATs in addition to cash incentives to encourage PA; results at 6 months were promising, but these effects were not sustained 6 months after the incentives were discontinued. Another recent study compared a standard weight loss intervention with a WAT-enhanced intervention and found the addition of WATs resulted in less weight loss over a 24-month period. Preliminary results from U.S. Army Performance Triad initiatives, which included units receiving WATs, found that though self-monitoring behaviors increased, only 38% of soldiers met all of the outlined activity goals. Further, though WAT use was initially high, it decreased sharply to approximately 30% after 2 months.

Privacy Protection and Network Operational Security

Federal regulatory, global health, and commercial security guidance universally acknowledge vulnerabilities in personally identifiable information (PII) data security. WATs are susceptible to both user privacy and network vulnerabilities, as they require users to submit PII in order to access services. Collecting, transmitting, storing, and sharing PII and other personal data about users entails risk. U.S. Federal regulatory agencies have warned that some of the most popular mobile health and fitness applications put potentially sensitive consumer data at risk by sharing it with scores of third-party companies. All WATs, even those with no Global Positioning System, are vulnerable to location tracking. Network systems supporting WAT devices are proprietary, making operational security and privacy concerns difficult to mitigate or manage.

Conclusions

The literature generally supports that when WATs are used as part of a structured, behavioral intervention, education, or feedback program, short-term increases in activity among some populations (e.g., lower activity strata) are obtainable. Research has yet to provide sufficient evidence on sustained WAT use or prolonged benefits to health and physical fitness. Finally, the proprietary nature of available WATs makes inherent privacy and operational security issues outside of user control.

INTRODUCTION

Engaging in regular physical activity (PA) has numerous benefits. Exercise can help control weight, reduce risk of certain diseases and cancers, improve physical, physiological, and mental health, and even increase life expectancy (Centers for Disease Control and Prevention, 2015). Nevertheless, nearly one in four adults in the United States fails to achieve weekly PA recommendations (Harvey, Chastin, & Skelton, 2013) and over two thirds are now considered overweight or obese (body mass index [BMI] ≥25.0; Flegal, Carroll, Kit, & Ogden, 2012). Members of the U.S. Armed Forces are not immune to these challenges; indeed, overweight and obesity rates among active duty service members have come to resemble those of the general population (Reyes-Guzman, Bray, Forman-Hoffman, & Williams, 2015; Smith et al., 2012). Given the military's history of stringent physical fitness requirements, these findings have implications for the recruitment, health, readiness, and retention of the warfighter.

In 2013, the U.S. Defense Health Board (DHB) reported on military obesity trends and the overall health and fitness of the Force. In their conclusions, they recommended the Department of Defense "develop and promote technology-based approaches to improved fitness" (p. 5). In addition, the DHB suggested interventions targeting military health and fitness should "incorporate exercise, healthy eating information, good sleep hygiene, behavioral modification, self-monitoring, relapse prevention, and structured follow-up by trained personnel" (p. 5).

Two years after the DHB report was published, the U.S. Navy announced plans to revise its Physical Readiness Program (Chief of Naval Operations, 2015). Objectives of this initiative included a desire to provide "a more realistic measure of health, fitness, and mission readiness [by incorporating] methods of assessing sleep patterns, activity, nutrition, and genetic risk factors" (para. 5). Proposed efforts toward achieving this end included new diet programs, changes to the Physical Readiness Test, and, most relevant to the current report, the conduct of "wearable-fitness device studies to monitor physical output and rest" (para. 4).

Wearable Fitness Devices

For years, research-grade accelerometers provided the gold standard for measuring PA. Accelerometers detect body movements over time, yielding data on the frequency, intensity, and duration of PA sessions. Through use of mathematical modeling, these data can be transformed into estimates of energy expenditure (EE), which can further be used to classify PA levels (e.g., low, moderate, or vigorous). Research-grade accelerometers, however, are costly, require technical knowledge and statistical expertise, and are limited to measuring ambulatory activity.

Over the last decade, a number of commercially available wearable activity trackers (WATs) have appeared on the market; Table 1 provides a list of WATs available at the time of this writing. These devices have become increasingly popular for their purported ability to provide users with real-time, tailored information about their daily activities, health, and personal fitness. As with research-grade devices, WATs use accelerometers, along with proprietary sensor technologies and data processing formulas, to generate and display data on movement, EE, and time spent in PA. These devices are available off-the-shelf, require limited training to operate,

and come at a fraction of the cost of gold standard devices; thus, WATs offer a practical alternative for everyday measurement of PA.

Independent research on commercially available WATs has been largely outpaced by updates in the WAT marketplace. Indeed, the average WAT life span is around nine months. Therefore, by the time any WAT can be scientifically evaluated and the results published, manufacturers are likely to develop and launch their next generation devices, including advancements to scoring algorithms and updates to software applications. Still, even with this lag, an increasing number of WAT studies have appeared in the literature addressing the benefits and shortcomings of these devices.

Reliability and Validity

The existing body of WAT research largely examines the reliability and validity of these devices; that is, how consistently and accurately these devices measure what they purport to measure. Reliability is commonly assessed by comparing data from a single WAT across multiple trials (i.e., test-retest or intra-device reliability) or by comparing multiple units of the same device worn simultaneously (inter-device reliability). Validity, on the other hand, is typically assessed by comparing WAT-derived measurements and similar metrics produced by gold standard devices. The most common output or data metrics evaluated in WAT research include step count, heart rate (HR), EE, and time spent in activity. However, with recent updates to WAT technologies, metrics on sleep duration and quality have begun to appear in the literature.

Sleep Assessments

Sleep plays an integral role in physiological (Knutson, Spiegel, Penev, & Van Cauter, 2007) and mental health (Breslau, Roth, Rosenthal, & Andreski, 1996; Smith et al., 2008), and adequate sleep is necessary for achieving high-level performance (Banks & Dinges, 2007). The gold standard for measuring sleep in research and clinical practice is polysomnography (PSG), which involves measurement of brain and muscle electrical activity. Unfortunately, this technique is costly, requires specialized training to conduct studies and interpret the data, and can be disruptive to sleep. Wrist-worn accelerometers, including those found in WATs, can serve as a proxy for estimating sleep by measuring body motion. The resulting data can be further refined through use of sleep logs, which inform sleep tracking algorithms of specific time periods during which sleep was attempted. However, unlike with PSG, this less invasive form of measurement produces sleep summary statistics only (e.g., time in bed, total sleep time [TST], sleep onset latency [SOL], wake after sleep onset [WASO], and sleep efficiency [SE]). Nevertheless, given the importance of sleep to health and well-being, it is likely that WATs will continue to incorporate sleep measures among their battery of assessments.

Behavior Change

Of course, for WATs to truly have value in health promotion activities, they need to go beyond reporting metrics. To be effective, WATs need to produce positive and lasting changes in health and PA behaviors. In 2014, the Telemedicine and Advanced Technology Research Center of the U.S. Army Medical Research and Materiel Command convened a workshop on "Leveraging Technology: Creating and Sustaining Changes for Health" (Teyhen et al., 2014). During this

event, military, academic, and industry participants explored best practices and research gaps in the use of WATs for promoting positive health outcomes. Though device accuracy and precision were listed as fundamental requirements, the panel determined "if the primary goal of the technology is *behavior change*, a 'good enough' accuracy threshold may exist."

Objectives

Commercially-available WATs are portrayed as user-friendly and beneficial to well-being, which raises the question of their potential application among military members. A growing number of today's service members are overweight and exhibit suboptimal PA and sleep behaviors (DHB, 2013; Barlas, Higgins, Pflieger, & Diecker, 2013). Consequently, the Navy has initiated efforts to revise its Physical Readiness Program to include plans to study WATs for measuring physical output and rest. In response to these initiatives, the current report was developed to summarize extant WAT research related to the Navy's objectives and of relevance to military populations. Specifically, this report:

- Examines evidence on the reliability and validity of WATs for measuring PA;
- Evaluates research on WATs for measuring sleep and sleep health;
- Summarizes literature on WATs in relation to behavior change; and
- Highlights privacy concerns and other issues of military interest.

METHODS

Separate search strategies were developed for each content area (i.e., reliability and validity, sleep, and behavior modification) under the guidance of a trained reference librarian. Specific search criteria are presented in Table 2. Articles were collected from PubMed, Ovid, Scopus, and Google Scholar and were limited to studies published in English language journals from January 2009 to July 2016. Only studies of commercially available devices were included. Studies unrelated to PA (or sleep) or that focused on specific clinical populations were excluded. In total, the searches identified 317 full-length studies. After reviewing for inclusion and exclusion criteria, 34 studies were retained for review and analysis.

RESULTS

Reliability and Validity

The literature search on WAT reliability and validity yielded 150 articles, of which 15 were retained. Key characteristics of these studies, including demographics, are summarized in Table 3. The majority of these studies were conducted in a controlled laboratory environment (11 studies) versus a field setting (four studies). Participants tended to be described as healthy or free of notable health conditions. The most commonly reported metrics included step count, HR, EE, and time spent in PA. Because these studies often used different testing and data reporting methodologies, the ability to make cross-study comparisons was limited.

Reliability

Only three studies provided data on WAT reliability; of these, all three reported on step count and only one reported on EE. No studies provide reliability data on HR or time spent in PA. Reliability tended to be assessed using correlational analyses. Criteria for interpreting the strength of the correlations varied across studies; however, Kooiman et al. (2015) recommended using the following cutoffs: >.90 (excellent), .75–.90 (good), .60–.75 (moderate), and <.60 (low). Of the three studies that provided reliability data for step count, one provided intra-device reliability estimates (Kooiman et al., 2015), one reported on inter-device reliability (Diaz et al., 2015), and the third reported on both types of reliability (O'Connell et al., 2016). Overall, reliability estimates for step count varied, with correlation values ranging from .53 to .90 across studies. These studies are summarized below.

In Kooiman et al.'s (2015) study, 33 participants wore three different WATs during two 30-minute sessions of treadmill walking. Test-retest reliability, assessed via intraclass correlation coefficient (ICC), rated good for two of the devices (ICC = .81 and .83) and poor for the third (ICC = .53).

Diaz and colleagues (2015) also used a treadmill design, though their study included a single test session featuring four 6-minute speed stages (ranging from 1.9 mph to 5.2 mph). Twenty-three participants simultaneously wore a unit of the same brand of WAT on each wrist, and the resulting correlation for step count was excellent (r = .90). Diaz et al. (2015) similarly calculated reliability for EE, with excellent results (r = .95).

In their field study, O'Connell et al. (2016) had 15 participants walk a prescribed path consisting of multiple surface types (e.g., grass, gravel, and stairs) for two test sessions, once wearing running shoes and once wearing hard-soled shoes. The path was nearly 2.5 miles in length (3,970 meters, plus 98 stairs) and, on average, took over 10,000 steps to complete. Prior to the study, the research team assessed WAT inter-device reliability by having an investigator wear multiple units simultaneously while walking 400 meters across a flat surface. Results of this process revealed a 13.67% variation in step count across devices. However, in the actual study, there were no statistically significant differences in step count between trials as recorded by the same device (correlation data were not provided). Thus, this study provided evidence for intra-device reliability, but not inter-device reliability, for the same WAT.

Validity

Across all 15 WAT validation studies, validity was assessed via comparison with gold standard measures. The most commonly investigated metric was step count (12 studies), followed by EE (eight studies), HR (two studies), and time spent in PA (two studies). As with the studies on reliability, studies reporting on WAT validity used a variety of statistical procedures and reporting methodologies. Analytic approaches included calculation of the absolute difference, percent agreement, correlation, and mean absolute percentage error (MAPE) between measured (WAT derived) and actual (gold standard) values. Table 4 summarizes the 15 WAT validation studies. Additional details, organized by data metric, are provided in the sections that follow.

Step Count. Step count validity studies included comparisons with observed step counts and/or step counts recorded by a research-grade device (e.g., accelerometer). Validity evidence varied by study design (treadmill vs. free-living), duration (minutes vs. days), pace (walking vs.

running), device placement (wrist vs. hip-worn), and, naturally, by the device itself. This lack of standardization limited the degree to which study results could be summarized in a cohesive manner, as well as the ability to generalize from the findings. For instance, across all step count studies, correlations between WAT derived and actual step count ranged from .67 to .97. Key differences in how these studies were conducted influenced the interpretation of the results, as detailed below.

In one study (Wallen, Gomersall, Keating, Wisløff, & Coombes, 2016), three WATs were evaluated while 22 participants completed a 1-hour protocol that included supine and seated rest, walking and running on a treadmill, and cycling on an ergometer. All three WATs slightly underestimated actual step count by 4–6% and correlations with the gold standard were moderate to good (r = .67 to .88). In another study (Ferguson, Rowlands, Olds, & Maher, 2015), 21 participants also wore multiple WATs, though in this case for 48 hours of free-living activity. Results again found the WATs to underestimate step count, one by −1,054 steps (range of difference: −4,693 to +1,804) and the other by −251 steps (range of difference: −1,978 to +2,252). Although correlations between the WAT-derived and actual step counts were excellent for these devices (r = .94 and .97, respectively), the large ranges of difference suggest both WATs were considerably less accurate in some trials than in others.

Lab-based studies that used a treadmill or prescribed path tended to yield encouraging step count validity estimates. Across eight laboratory studies investigating 19 different WATs, 10 devices were over 95% accurate while only four were less than 90% accurate. Among these studies, the type of movement performed influenced step count accuracy. For instance, Nelson, Kaminsky, Dickin, and Montoye (2016) reported greater step count accuracy for ambulatory tasks versus sedentary or general household tasks, and Chen and colleagues (2016) found accuracy improved as participants moved at faster speeds.

For field-based studies, step count validity varied considerably. Across four studies investigating eight different WATs, MAPE ranged from 1.4% to 29.0%. Among these studies, step count accuracy appeared to be device- and task-dependent. For example, in Kooiman et al.'s (2015) free-living condition, 56 participants simultaneously wore three different WATs while performing 7.5 hours of normal daily activities. Two of the devices were over 95% accurate, while the third miscounted steps by 24%. Alternatively, in their investigation on nine surface types, O'Connell and colleagues (2016) found step count accuracy suffered when particpants ascended or descended stairs (MAPE = 11.33% and 6.48%, respectively) but was otherwise 95% accurate or better.

Energy Expenditure. For validity studies on EE, the gold standard measure was kilocalories (kcal) assessed via indirect calorimetry or accelerometer. Correlations to this reference varied widely (r = .16 to .86), as did MAPE (11.2% to 53.0%). More often than not, WATs underestimated EE. However, there were some exceptions to this finding.

Wallen et al. (2016) had participants engage in diverse activities over 1 hour, and all three devices in their study underestimated EE (range of difference: −26.1 to −224.6 kcal). However, only one in four WATs did the same in Bai et al.'s (2016) study (mean difference: −42.3 kcal), while two others overestimated EE (range of difference: 20.4 to 72.4 kcal), and the fourth failed

to register complete data due to a syncing error. Further, in Dondzila and Garner's (2016) treadmill study, the WAT overestimated EE at the slowest speed but underestimated EE at three faster speeds. Thus, accuracy of EE measurement varied both across and within devices, depending on the activity.

Ferguson et al.'s (2015) 48-hour free-living study further revealed the range of error associated with EE measurement. Both WATs studied correlated similarly with the research grade device (r = .74 and .79) and each was found to underestimate the criterion, one by −866 kcal (range of difference: −1,937 to −94) and the other by −468 kcal (range of difference: −996 to 108). These findings resembled Ferguson et al.'s step count data, with the large ranges of difference indicating both WATs were less accurate in some trials than in others. Of note, the WATs evaluated in this study had opposite strengths, in that the device that performed better for step count performed less well for EE.

Heart Rate. Newer WATs have started to include embedded monitors that measure HR via photoplethysmography. The two studies reviewed herein included HR validity data across five total devices, with mostly promising results. In one study, participants walked 200, 500, and 1,000 steps on a treadmill while wearing two different WATs (El-Amrawy & Nounou, 2015). Compared with the gold standard (clinical pulse oximeter), both devices were found to be highly accurate (percent agreement: 92.5% to 99.9%). In another study, participants performed sedentary, ambulatory, and cycling tasks while wearing three different WATs (Wallen et al., 2016). Correlations for HR between these devices and a reference device were moderate, good, and excellent (r = .67, .81, and .95, respectively).

Time in Physical Activity. Two validity studies on four total devices examined time spent in moderate to vigorous physical activity (MVPA; Ferguson et al., 2015; Rosenberger, Buman, Haskell, McConnell, & Carstensen, 2016). Both studies used free-living activity designs but with different data reporting methods. In Ferguson et al.'s (2015) 48-hour study, correlations between the WAT estimates of MVPA and the reference device were good (r = .81 and .79). One device overestimated time in MVPA by 18 minutes (range of difference: −4.7 to 96.5 minutes), while the other underestimated the criterion by 15.2 minutes (range of difference: −79.7 to 36.3 minutes). Results from Rosenberger et al.'s (2016) 24-hour study were less promising, with one device miscalculating time in MVPA by 48 minutes (MAPE = 52%) and the other by 598 minutes (MAPE = 92%). These studies indicated that WAT measurements of time in PA may be inconclusive at best.

Sleep Assessments

The literature search on WAT use for sleep measurement yielded 140 articles, of which 15 were retained. Of these 15 studies, eight were original research papers and seven were literature reviews. Key characteristics of these studies are summarized in Table 5, including sample size, measures tested, and results. All studies occurred over a single data collection night; therefore, the reproducibility of sleep outcomes across several nights could not be examined.

Six of eight studies found WATs overestimated TST (range: 8.0 to 67.1 minutes), five of six studies found overestimations in SE (range: 1.8–14.5%), and four of five studies found WATs

underestimated WASO (range: 5.6 to 32.2 minutes). The findings indicated the majority of WATs were not effective in discriminating between sleep and wakefulness due to misidentification of periods of wakefulness as sleep, which is concerning since an overnight sleep period may involve a considerable amount of wakefulness, especially in clinical populations (e.g., insomnia and other anxiety-related disorders) and conditions and settings conducive to sleep disruption (e.g., environmental noise, children, and daytime sleep with shift work).

One study provided evidence that WATs could be equivalent to, or even better than, an accepted standard (Toon et al., 2016). This study was conducted with a specific clinical population of children and adolescents who were varied in age and developmental stages, which may have resulted in these positive findings. Compared with PSG, WAT sensitivity and specificity for sleep were equivalent to a wrist-worn gold standard measure, and the WAT performed better than this reference device at estimating SOL. However, as the age of participants increased, WAT performance decreased, presumably due to differences in wrist activity movements during wakefulness across age (younger children tended to have larger arm/wrist movements that may have triggered algorithm thresholds for wakefulness compared with adolescents or adults).

Three additional studies became available after the conduct of the larger literature review. In one, Mantua, Gravel, and Spencer (2016) evaluated four WATs against PSG in 40 healthy, young adults. Three of the devices measured sleep stages (i.e., classifying light vs. deep sleep), with mixed results. However, since no sensitivity analyses were performed, the contribution of these results to the greater validation effort was minimal. Correlations with reported TST were good for all devices ($r > .75$); however, none of the devices correlated significantly with SE. Similar results were reported by de Zambotti et al. (2016) in their evaluation of 32 healthy adolescents. Results again indicated sensitivity for sleep was high (.97) but specificity was low (.42). In a third study, Rosenberger and colleagues (2016) compared sleep metrics in 40 young adults using two WATs and the Zmachine, a three-lead electroencephalography system with prior validation. Both WATs performed poorly, overestimating TST. Once again, sensitivity and specificity values were not reported in this study, which further highlighted the lack of standardization in WAT research studies.

Behavior Change

The literature search on WAT-based behavior change studies identified 99 studies, of which 10 were retained. These included studies on the prevalence of behavior change techniques (BCTs) in commercially available WATs, research on WAT-based behavior change interventions, and surveys on the behavior-based outcomes of users' personal WAT experiences. Results of these studies are summarized in Table 6 and the sections that follow. Unpublished preliminary results from the U.S. Army Performance Triad (P3) initiatives, which included military units receiving WATs, are also discussed in this section.

Behavior Change Techniques

Michie and colleagues (Abraham & Michie, 2008; Michie et al., 2011; Michie et al., 2013) have identified, defined, and developed taxonomies for a number of BCTs used in behavior change interventions. Using Michie et al.'s (2013) taxonomy of 93 BCTs, Lyons, Lewis, Mayrsohn, and

Rowland (2014) found all 13 WATs reviewed included features for self-monitoring, feedback, environmental change (i.e., adding the WAT to the user's environment), goal setting, and an emphasis on the discrepancy between current behavior and goal behavior. In addition, over half of the devices provided updates of behavioral goals, social support, social comparison, prompts/cues, rewards, and a focus on past successes. These authors were careful to note, however, that the prevalence of BCTs may not be as important as the effective application of the techniques; specifically, "a system with fewer but more effective techniques may ultimately produce a greater impact than a system with more numerous but less effective ones" (Lyons et al., 2014).

In 2016, Mercer, Li, Giangregorio, Burns, and Grindrod further examined BCTs in WATs, this time using Michie et al.'s (2011) 40-item taxonomy specifically designed for PA and healthy eating interventions. Their analyses found all seven WATs reviewed had the same nine features, i.e., they all (1) prompted users to self-monitor their activity levels, (2) review goals, and (3) focus on past successes, while providing (4) feedback on performance, (5) rewards contingent upon successful behavior, (6) normative information on others' behavior, and (7) the ability to see others' approval, as needed to (8) facilitate social comparison and (9) plan social support/social change. The authors noted most of these BCTs, including goal setting, feedback, and social support, have been well documented as effective among younger adults (French, Olander, Chisholm, & Mc Sharry, 2014), which may explain their universal presence in modern WATs.

Behavior Change Interventions

Beyond determining BCT presence in WATs, a practical consideration is the degree to which these BCTs actually produce meaningful changes in behavior. Research on this issue has produced mixed results, though differences in methodologies make it challenging to determine whether the findings were based on the device or the design. For instance, Groendal, Vasold, Knous, and Schlaff (2014) examined if a WAT-based PA intervention could increase activity in inactive adults. Participants who did not meet PA recommendations (n = 16) were provided a WAT to wear for 8 weeks, while a control group of physically active participants (n = 17) was told to maintain current activity levels and was not provided a device. After the 8-week trial, the intervention participants significantly increased their self-reported PA (baseline = 74 minutes; week 8 = 216 minutes; $t = -3.545$, $p = .002$), while the control group reported modest but nonsignificant gains (baseline = 358 minutes; week 8 = 396 minutes). Of note, at a 6-month follow-up, the treatment group maintained elevated activity levels (205 minutes), despite not wearing the WAT since week 8.

Valbuena, Miltenberger, and Solley (2015) also investigated whether WATs could improve health behaviors of inactive individuals. In their three-phase study, seven overweight participants (BMI $\geq$25) were issued covered devices (data inaccessible) to assess baseline activity levels (phase one). Once baseline measurements were achieved, the WATs were uncovered (data made accessible) and measurements of changes in activity level were obtained until they stabilized (phase two). In the final phase, participants maintained use of their WATs and received individualized behavioral coaching in the form of a phone call from the research staff. This coaching included tailored feedback and social support (praise), as well as recommendations on how participants might improve their fitness. Results suggested having access to the WAT data

alone (phase two) was sufficient to increase the PA levels for three of the seven participants. With the subsequent addition of the behavioral coach, all six participants (one dropped out) demonstrated substantive PA gains, with step count increases ranging from 10.2% to 89.1% over baseline. Despite having a small sample, the extended duration of this study (range: 147 to 449 days) provided a rare look at the lasting effects of WAT use on activity levels, including benefits received from adding BCTs (behavioral coaching) beyond those provided by the technology itself.

Similar results were observed by Ross and Wing (2016), who found overweight participants who received a WAT and structured phone calls (calls on general weight loss techniques, such as goal setting, problem solving, stimulus control, social support, and relapse prevention) lost more weight (mean change ± SE = −6.40 ± 1.17 kg; −7.37 ± 1.29%) over a 6-month intervention period than did those who received a WAT or pedometer alone (−4.04 ± 1.37 kg; −4.35 ± 1.29%).

In a related study, Wang et al. (2015) examined the effects of a WAT plus direct contact intervention on the fitness of overweight and obese adults using text messaging. In their 6-week study, 67 participants were randomly assigned to either receive a WAT or a WAT plus three daily text messages prompting engagement in PA. Unlike in the aforementioned studies, this control group (WAT only) exhibited a significant increase in PA, while the WAT + Text group did not. Follow-up analyses revealed that text messaging was effective for 1 week but was insufficient thereafter to promote any change. The authors posited the automated nature of the texts, specifically their lack of individualized tailoring, may have been the issue. Indeed, many participants reported they stopped reading the texts after a few days once they realized the messages were not tied to their own performance. The three texts per day reminders were also deemed to be too frequent to be helpful.

One study examined the effectiveness of WAT alone or in combination with cash incentives or charitable donations during a 6-month intervention (Finkelstein et al., 2016). Participants included 800 desk-bound office workers, over half of whom were considered overweight or obese. The cash incentive was most effective at increasing MVPA at the end of the 6-month intervention; however, this effect was not sustained 6 months after the incentives were discontinued. At the 12-month follow-up, there was no evidence of improvements in health outcomes using WATs, regardless of whether incentives were in place.

Another study examined a 24-month WAT-enhanced weight loss intervention in young overweight or obese adults (Jakicic et al., 2016). Participants were placed on a low calorie diet and were asked to increase PA and attend group counseling sessions. At 6 months, all participants received telephone counseling sessions, text message prompts, and access to study materials on a website. The control group started self-monitoring their diet and PA using a website, while the intervention group was provided a WAT and web interface to monitor diet and PA. Although both groups experienced comparable improvements in body composition, fitness, PA, and diet, the control group lost more weight than the WAT-intervention group. These findings, again, suggested that WATs may not provide increased benefits beyond those achieved through other weight loss approaches.

A recently completed pilot study of the Army P3 program incorporated WATs to try and enhance individual and unit performance through improvements in sleep, activity, and nutrition. During the 26-week study, 2,200 soldiers used WATs to track their PA against three recommended fitness goals: 10,000 steps/day, at least 150 minutes of moderate aerobic exercise per week, and at least 2 days of resistance training per week (Lilley, 2015). Preliminary unpublished results showed although soldiers initially used the WAT regularly, usage significantly decreased to about 30% after 8 weeks into the study (T. Brahmbhatt, personal communication, February 18, 2016). Furthermore, despite mandatory unit physical training, only 29–42% of participating soldiers met all of the outlined activity goals.

Behavior Change Experiences

Two studies examined users' personal experiences with WATs through open-ended questioning. In one (Karapanos, Gouveia, Hassenzahl, & Forlizzi, 2016), the researchers surveyed 133 WAT users on their most memorable WAT experiences. Responses were analyzed according to a need fulfillment framework, and a two-factor structure was revealed. Specifically, WATs fulfilled users' (1) need for *physical thriving* (feelings of competence and self-esteem, achieved through self-monitoring, goal setting, and feedback features), and (2) need for *relatedness* (social comparison and popularity, achieved through data sharing, competition, and online communities). However, not all users' experiences were the same, as the data revealed two classes: *purposive* and *explorative* users. Purposive users acquired their WATs deliberately to achieve a healthier lifestyle, to quantify and track their PA, or to overcome barriers to exercise. Explorative users received their WATs as gifts, purchased them impulsively, or purchased them to support friends and family in their efforts to achieve healthier lifestyles. Of these groups, purposive users were more likely to be engaged with their WATs and maintain use of their WATs over time. As this study suggested, behavior change is not only a function of the device's features but also of individual differences in the user population. In this case, the users' motivations and self-set goals prior to using their WATs affected their overall experiences and likelihood of prolonged use.

In a second open-ended study, Chang, Lu, Yang, and Luarn (2016) found most WAT users purchase their devices deliberately for tracking their PA and progress toward health goals. This is to say, the device itself was not viewed as the primary driver of performance but rather as a means to an end (WATs provided more awareness of activity than motivation to perform). Participants reported they relied on social support and social networking services to remain engaged with their WATs; they also expressed appreciation for team-based challenges that inspired users to undertake group missions in order to reach shared goals. However, some users directly reported privacy concerns regarding online postings and/or sharing of their PA levels. Further, only one in four respondents expressed interest in sharing their exercise information through social media. These findings provided insight on how inclusion of certain BCTs (social comparison) may be valued by some users but rejected by others.

DISCUSSION

This report provided an overview of research on WAT reliability and validity, sleep assessment, and behavior change. The literature search was restricted to WATs that were currently available

on the market, and studies were excluded if they focused on clinical populations (among other exclusions; see Table 2).

Reliability and Validity

Fifteen studies were reviewed on WAT reliability and validity. Reliability studies included test-retest studies as well as inter-device reliability studies. Validity studies included comparisons with gold standard measures of step count, HR, time spent in PA, and EE. Subjects participated in controlled treadmill tasks at various speeds, structured activities at various intensity levels, and unstructured free-living activities for various time durations.

Test-retest and inter-device reliability tended to be good to excellent for most devices across all tasks. This finding suggests that if a single brand/type of device was used by all subjects undergoing the same tasking, similar results could be expected. However, it does not suggest that these findings would be accurate.

Validity was assessed most often through correlations and calculation of MAPE. Results varied tremendously across all devices and tasks, and no device was accurate across all activity types. WATs were most accurate for detecting step count, though results varied by task and device. Hip-worn WATs were more accurate in detecting step count than wrist-worn devices. WATs tended to more accurately measure step count during treadmill tasks at high speeds. They were less accurate at slow speeds or with less structured tasks. Cycling tended to be the most challenging activity for WATs to accurately measure in terms of step count and EE. Studies that involved multiple activities demonstrated a high degree of variability in measurement accuracy, with some devices underestimating criterion values and others overestimating criterion values for any given task. Free living and structured activity studies that lasted several hours (or days) demonstrated that averaging WAT values across multiple activities may present misleading measures of accuracy, since the under- and overestimated activity values tend to cancel each other out (Lee, Kim, & Welk, 2014).

Based on this review, WAT researchers are encouraged to select the device(s) to study based on the specific activities of interest. The device should be calibrated before the research begins using multiple gold standard measures and under multiple intensity levels. The study samples should also be representative of the target population in age, sex, and BMI. Furthermore, it is recommended that future WAT research adopt a standardized method for reporting data. This practice would enhance cross-study comparisons and enable improved generalization of results.

Sleep Assessments

Seven review papers and eight original papers were reviewed based on WAT sleep assessments. All studies took place over a single night, which precluded night-to-night variability assessment. For validity, sensitivity was very high for the agreement of sleep epochs between WATs and gold standards, but specificity was low for the agreement of wake epochs. WATs tended to overestimate TST, underestimate WASO, and overestimate SE.

Due to a lack of available research, the accuracy and reliability of sleep information from WATs is still mostly unknown. The misidentification of periods of wake as sleep was most commonly reported as a barrier to their accuracy and therefore usefulness, especially for individuals who experienced more disrupted sleep. Thus, the consensus among researchers is that currently available WATs show theoretical promise but do not perform well enough for adoption as health and medical monitoring tools for accurately measuring sleep.

One possible application of currently available WATs would be to support sleep therapy interventions. Cognitive and behavioral interventions for insomnia have been shown to be effective for improving disrupted sleep (Qaseem et al., 2016; Troxel, Germain, & Buysse, 2012), and the educational content of these interventions could be integrated into WAT software platforms. WAT data could also be used to identify behaviors that have an impact on sleep. For example, exercising too close to bedtime can delay sleep onset (Oda & Shirakawa, 2014). WAT data could also contribute to recommendations for improving sleep and next-day alertness by providing bed and wake time suggestions based on the prior night's sleep data, as consistent sleep schedules are important for sleep quality (Borbély & Achermann, 1999; Manber, Bootzin, Acebo, & Carskadon, 1996). WATs that collect additional user-reported information, such as perceived sleep quality and engagement in common behaviors known to affect sleep, may be better positioned to accomplish this goal. Linking habits and user perception of sleep quality to objective sleep data, even if those data are not entirely accurate, can result in greater personal awareness of how disruptive behaviors impact sleep. This is an area that will need to be explored with longitudinal studies in a variety of settings and populations.

Behavior Change

Ten studies examining WAT use related to behavior change were reviewed. The results of these studies were effectively summarized by the "Leveraging Technology" workshop participants who issued the following guidance (Teyhen et al., 2014):

- Not all individuals respond to PA interventions in the same manner; some may benefit from less intensive interventions, whereas others may require more.
- WATs should provide both strategies and incentives to help individuals reach personal health goals. Feedback should be personalized and adaptable based on an individual's preferences, readiness to change, and pace of goal achievement.
- To be effective in the long term, WAT-based fitness interventions should be made enjoyable (i.e., emphasize fun through gamification, competition, incentives, and social support).

Whether WATs can produce lasting behavior change in healthy populations remains unanswered. Surveys report that more than half of U.S. consumers who have owned a modern WAT no longer use it (Ledger & McCaffrey, 2014). In order to provoke sustained long-term use of WATs, the research suggests device interventions need to move beyond presenting data (steps, calorie count) and begin to leverage behavioral economic and social psychology theories that contribute to habit formation, social motivation, and goal achievement.

Research on BCTs in wearable technologies suggests PA interventions that incorporate WATs succeed in motivating participants to initiate increased activity; however, most of these studies

were conducted with unhealthy, overweight, or elderly populations rather than demographics that are representative of active duty service members. Evidence suggests that with proper preparation, support, and feedback, WATs can serve as useful tools for assisting individuals in maintaining a healthy lifestyle. However, to increase the likelihood of WAT program success by modifying a person's long-term health-related behavior likely requires more than provision of a fitness tracking device. A systematic approach that considers the unique characteristics of the device, the individual, and the environment may also be necessary.

Behavior-based research gaps that are of particular relevance to military populations involve understanding habit management and social networking. Understanding how contextual changes and times of instability affect habit management is especially important because service members undergo military moves, deployments, and retirement. Social networking may support positive habit formation by providing a stable context that can survive these disruptions through a lasting virtual community. Future WAT studies need to be conducted with active duty personnel to determine whether service members could benefit from a mobile platform aimed at improving overall health.

Privacy Protection and Network Operational Security

If WAT use is to be promoted among military members, the risks and implications of transmitting, storing, and sharing WAT-collected health data must be considered. The Federal Trade Commission (2016) has warned that some of the most popular mobile health and fitness applications put potentially sensitive consumer data at risk by sharing it with scores of third-party companies. Information that could easily be traced back to individuals included everything from eating habits to medical symptom searches to walking/running routes.

For myriad reasons, most WATs do not have a user interface for direct data analysis and display. Instead, data must be transferred to a processing center where it can be aggregated and interpreted before being presented in a user interface. Consequently, the life cycle of data handled by most WAT systems involves three stages: a data collection phase, a transmission phase, and an Internet-based cloud storage and analysis phase with potential feedback loop. This "ecosystem" entails three main areas in which data could be at risk: on the device (storage), in transit (transmission), and in the cloud (storage).

Never before has such a vast amount of information been collected, transmitted, and stored about users. People are freely and actively engaging in the collection of information about themselves, and how these data are managed in the cloud is generally outside of the control or visibility of the users. Traditional personally identifiable information (PII) can tell somebody about who we are, where we live, and how to contact us. Additional information generated by self-tracking services can tell somebody about what we do, where we are or have been, and when and potentially why we are doing something. When additional self-tracking information is combined with traditional PII, the potential for abuse becomes even greater. As data are aggregated and relationships between data are formed, the resulting insight can be used to predict the future behavior of people, which is beneficial to marketing but worrisome as it relates to privacy and security.

Profiling

Self-tracking service providers have quickly caught on to other potential business applications for data generated by self-tracking technologies. Many organizations use profiling to target, exclude, or even discriminate against certain types of people based on personal information collected. Details provided by users to self-tracking services could enable marketers to organize and target certain types of users. Profiling is of concern to privacy and human rights advocates because it can be easily misused to the disadvantage of certain groups or minorities.

User Location or Stalking

Security risks have been identified in a large number of self-tracking devices and applications. All WATs, even those that do not feature a Global Positioning System (GPS), are vulnerable to location tracking (Barcena, Wueest, & Lau, 2014). Accurate and real-time location-based tracking can be useful for some activities; however, location-based self-tracking PII could be abused for criminal purposes. Figure 1 provides a screenshot from a GPS-enabled WAT application. Exercise routines and probable location could be derived from this information, which, again, is not only stored on the device but remains indefinitely in the cloud.

CONCLUSIONS

The current review was conducted to examine the scientific evidence on WAT validity and reliability, with particular attention to the military readiness context. Of the studies reviewed, most indicated that the utility and accuracy of WATs is highly variable, and few focused on reliability and validity in a relevant setting. This, combined with attempting to evaluate a technology advancing at a more rapid rate than one can study, makes it impossible to conclude definitively on the value of current WATs. The literature does suggest that WATs accurately document step count and are getting better at assessing EE; however, they remain unreliable for sleep measurement. The literature also suggests that WATs are most effective in having an impact on health and fitness when they are part of a behavior modification program, especially when used among overweight populations, and they tend to do well in motivating participants to initiate increased activity in the short term.

The question of whether WATs can produce lasting behavior change in healthy, physically fit populations remains unanswered. As the military and the Navy in particular address physical fitness, health, and readiness issues, including modernization of related programs, WATs could play a role. However, any incorporation of WATs into health and fitness programs should involve validation, rigorous scientific evaluation, and take into account the unique social and security concerns of the military context.

REFERENCES

Abraham, C., & Michie, S. (2008). A taxonomy of behavior change techniques used in interventions. *Health Psychology, 27*(3), 379–387. doi:10.1037/0278-6133.27.3.379

Bai, Y., Welk, G. J., Nam, Y. H., Lee, J. A., Lee, J. M., Kim, Y., . . . Dixon, P. M. (2016). Comparison of consumer and research monitors under semistructured settings. *Medicine and Science in Sports and Exercise, 48*(1), 151–158. doi:10.1249/MSS.0000000000000727

Banks, S., & Dinges, D. F. (2007). Behavioral and physiological consequences of sleep restriction. *Journal of Clinical Sleep Medicine, 3*(5), 519–528.

Barcena, M. B., Wueest, C., & Lau, H. (2014). Security response: How safe is your quantified self? Retrieved from https://www.symantec.com/content/dam/symantec/docs/white-papers/how-safe-is-your-quantified-self-en.pdf

Barlas, F. M., Higgins, W. B., Pflieger, J. C., & Diecker, K. (2013). *2011 Health Related Behaviors Survey of Active Duty Military Personnel*. Retrieved from http://www.dtic.mil/get-tr-doc/pdf?AD=ADA582287

Borbély, A. A., & Achermann, P. (1999). Sleep homeostasis and models of sleep regulation. *Journal of Biological Ryhythms, 14*(6), 557–568.

Breslau, N., Roth, T., Rosenthal, L., & Andreski, P. (1996). Sleep disturbance and psychiatric disorders: A longitudinal epidemiological study of young adults. *Biological Psychiatry, 39*(6), 411–418.

Brooks, G., Fahey, T., & Baldwin, K. (Eds.). (2004). *Exercise physiology: Human bioenergetics and its applications* (4th ed.). New York, NY: McGraw-Hill Education.

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. doi:10.1001/jama.2014.17841

Centers for Disease Control and Prevention. (2015). Physical activity and health. Retrieved from https://www.cdc.gov/physicalactivity/basics/pa-health/

Chang, R. C., Lu, H. P., Yang, P., & Luarn, P. (2016). Reciprocal reinforcement between wearable activity trackers and social network services in influencing physical activity behaviors. *Journal of Medical Internet Reserach mHealth and uHealth, 4*(3), e84. doi:10.2196/mhealth.5637

Chen, M. D., Kuo, C. C., Pellegrini, C. A., & Hsu, M. J. (2016). Accuracy of wristband activity monitors during ambulation and activities. *Medicine and Science in Sports and Exercise, 48*(10), 1942–1949. doi:10.1249/MSS.0000000000000984

Chief of Naval Operations. (2015). NAVADMIN 178/15: Physical readiness program policy changes. Retrieved from http://www.public.navy.mil/bupers-npc/reference/messages/Documents/NAVADMINS/NAV2015/NAV15178.txt

Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J., & Gillin, J. C. (1992). Automatic sleep/wake identification from wrist actigraphy. *Sleep, 15*(5), 461–469.

de Zambotti, M., Baker, F. C., & Colrain, I. M. (2015). Validation of sleep-tracking technology compared with polysomnography in adolescents. *Sleep, 38*(9), 1461–1468. doi:10.5665/sleep.4990

de Zambotti, M., Baker, F. C., Willoughby, A. R., Godino, J. G., Wing, D., Patrick, K., & Colrain, I. M. (2016). Measures of sleep and cardiac functioning during sleep using a

multi-sensory commercially-available wristband in adolescents. *Physiology and Behavior, 158*, 143–149. doi:10.1016/j.physbeh.2016.03.006

de Zambotti, M., Claudatos, S., Inkelis, S., Colrain, I. M., & Baker, F. C. (2015). Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiology International, 32*(7), 1024–1028. doi:10.3109/07420528.2015.1054395

Defense Health Board. (2013). Implication and trends in obesity and overweight for the Department of Defense. Retrieved from http://www.health.mil/Reference-Center/Reports/2013/11/22/DHB-Implications-of-Trends-in-Obesity-and-Overweight-for-the-DoD-Fit-to-fight-fit-for-life

Diaz, K. M., Krupka, D. J., Chang, M. J., Peacock, J., Ma, Y., Goldsmith, J., . . . Davidson, K. W. (2015). Fitbit®: An accurate and reliable device for wireless physical activity tracking. *International Journal of Cardiology, 185*, 138–140. doi:10.1016/j.ijcard.2015.03.038

Diaz, K. M., Krupka, D. J., Chang, M. J., Shaffer, J. A., Ma, Y., Goldsmith, J., . . . Davidson, K. W. (2016). Validation of the Fitbit One® for physical activity measurement at an upper torso attachment site. *BMC Research Notes, 9*, 213. doi:10.1186/s13104-016-2020-8

Dondzila, C., & Garner, D. (2016). Comparative accuracy of fitness tracking modalities in quantifying energy expenditure. *Journal of Medical Engineering and Technology, 40*(6), 325–329. doi:10.1080/03091902.2016.1197978

El-Amrawy, F., & Nounou, M. I. (2015). Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial? *Healthcare Informatics Research, 21*(4), 315–320. doi:10.4258/hir.2015.21.4.315

Federal Trade Commission. (2016). Mobile health app developers: FTC best practices. Retrieved from https://www.ftc.gov/tips-advice/business-center/guidance/mobile-health-app-developers-ftc-best-practices

Ferguson, T., Rowlands, A. V., Olds, T., & Maher, C. (2015). The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: A cross-sectional study. *International Journal of Behavioral Nutrition and Physical Activity, 12*, 42. doi:10.1186/s12966-015-0201-9

Finkelstein, E. A., Haaland, B. A., Bilger, M., Sahasranaman, A., Sloan, R. A., Nang, E. E., & Evenson, K. R. (2016). Effectiveness of activity trackers with and without incentives to increase physical activity (TRIPPA): A randomised controlled trial. *Lancet Diabetes and Endocrinology, 4*(12), 983–995. doi:10.1016/s2213-8587(16)30284-4

Flegal, K. M., Carroll, M. D., Kit, B. K., & Ogden, C. L. (2012). Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010. *Journal of the American Medical Association, 307*(5), 491–497. doi:10.1001/jama.2012.39

French, D. P., Olander, E. K., Chisholm, A., & Mc Sharry, J. (2014). Which behaviour change techniques are most effective at increasing older adults' self-efficacy and physical activity behaviour? A systematic review. *Annals of Behavioral Medicine, 48*(2), 225–234. doi:10.1007/s12160-014-9593-z

Groendal, E. G., Vasold, K. L., Knous, J. L., & Schlaff, R. A. (2014). Sustained effect of Fitbit technology on physical activity levels in inactive individuals [Abstract]. *Medicine and Science in Sports and Exercise, 46*(5S), 109.

Harvey, J. A., Chastin, S. F., & Skelton, D. A. (2013). Prevalence of sedentary behavior in older adults: A systematic review. *International Journal of Environmental Research and Public Health, 10*(12), 6645–6661. doi:10.3390/ijerph10126645

Jakicic, J. M., Davis, K. K., Rogers, R. J., King, W. C., Marcus, M. D., Helsel, D., . . . Belle, S. H. (2016). Effect of wearable technology combined with a lifestyle intervention on long-term weight loss: The IDEA randomized clinical trial. *Journal of the American Medical Association, 316*(11), 1161–1171. doi:10.1001/jama.2016.12858

Karapanos, E., Gouveia, R., Hassenzahl, M., & Forlizzi, J. (2016). Wellbeing in the making: Peoples' experiences with wearable activity trackers. *Psychology of Well-Being, 6*, 4. doi:10.1186/s13612-016-0042-6

Knutson, K. L., Spiegel, K., Penev, P., & Van Cauter, E. (2007). The metabolic consequences of sleep deprivation. *Sleep Medicine Reviews, 11*(3), 163–178. doi:10.1016/j.smrv.2007.01.002

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*, 24. doi:10.1186/s13102-015-0018-5

Ledger, D., & McCaffrey, D. (2014). Inside wearables: How the science of human behavior change offers the secret to long-term engagement. Retrieved from http://endeavourpartners.net/assets/Endeavour-Partners-Wearables-White-Paper-20141.pdf

Lee, J. M., Kim, Y., & Welk, G. J. (2014). Validity of consumer-based physical activity monitors. *Medicine and Science in Sports and Exercise, 46*(9), 1840–1848. doi:10.1249/MSS.0000000000000287

Lilley, K. (2015, July 27). 20,000 soldiers tapped for Army fitness program's 2nd trial. *Army Times*. Retrieved from https://www.armytimes.com/story/military/careers/army/2015/07/27/-performance-sleep-nutrition-fitness/70212286/

Lyons, E. J., Lewis, Z. H., Mayrsohn, B. G., & Rowland, J. L. (2014). Behavior change techniques implemented in electronic lifestyle activity monitors: A systematic content analysis. *Journal of Medical Internet Research, 16*(8), e192. doi:10.2196/jmir.3469

Manber, R., Bootzin, R. R., Acebo, C., & Carskadon, M. A. (1996). The effects of regularizing sleep-wake schedules on daytime sleepiness. *Sleep, 19*(5), 432–441.

Mantua, J., Gravel, N., & Spencer, R. M. (2016). Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography. *Sensors (Basel), 16*(5), 646. doi:10.3390/s16050646

Meltzer, L. J., Hiruma, L. S., Avis, K., Montgomery-Downs, H., & Valentin, J. (2015). Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. *Sleep, 38*(8), 1323–1330. doi:10.5665/sleep.4918

Mercer, K., Li, M., Giangregorio, L., Burns, C., & Grindrod, K. (2016). Behavior change techniques present in wearable activity trackers: A critical analysis. *Journal of Medical Internet Research mHealth and uHealth, 4*(2), e40.

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 and Health, 26*(11), 1479–1498. doi:10.1080/08870446.2010.540664

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., . . . 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. doi:10.1007/s12160-013-9486-6

Montgomery-Downs, H. E., Insana, S. P., & Bond, J. A. (2012). Movement toward a novel activity monitoring device. *Sleep and Breathing, 16*(3), 913–917. doi:10.1007/s11325-011-0585-y

Nelson, M. B., Kaminsky, L. A., Dickin, D. C., & Montoye, A. H. (2016). Validity of consumer-based physical activity monitors for specific activity types. *Medicine and Science in Sports and Exercise, 48*(8), 1619–1628. doi:10.1249/MSS.0000000000000933

O'Connell, S., ÓLaighnin, G., Kelly, L., Murphy, E., Beirne, S., Burke, N., . . . Quinlan, L. R. (2016). These shoes are made for walking: Sensitivity performance evaluation of commercial activity monitors under the expected conditions and circumstances required to achieve the international daily step goal of 10,000 steps. *PLoS One, 11*(5), e0154956. doi:10.1371/journal.pone.0154956

Oda, S., & Shirakawa, K. (2014). Sleep onset is disrupted following pre-sleep exercise that causes large physiological excitement at bedtime. *European Journal of Applied Physiology, 114*(9), 1789–1799. doi:10.1007/s00421-014-2873-2

Qaseem, A., Kansagara, D., Forciea, M. A., Cooke, M., Denberg, T. D., & Clinical Guidelines Committee of the American College of Physicians. (2016). Management of chronic insomnia disorder in adults: A clinical practice guideline from the American College of Physicians. *Annals of Internal Medicine, 165*(2), 125–133. doi:10.7326/M15-2175

Reyes-Guzman, C. M., Bray, R. M., Forman-Hoffman, V. L., & Williams, J. (2015). Overweight and obesity trends among active duty military personnel: A 13-year perspective. *American Journal of Preventive Medicine, 48*(2), 145–153. doi:10.1016/j.amepre.2014.08.033

Rosenberger, M. E., Buman, M. P., Haskell, W. L., McConnell, M. V., & Carstensen, L. L. (2016). Twenty-four hours of sleep, sedentary behavior, and physical activity with nine wearable devices. *Medicine and Science in Sports and Exercise, 48*(3), 457–465. doi:10.1249/MSS.0000000000000778

Ross, K. M., & Wing, R. R. (2016). Impact of newer self-monitoring technology and brief phone-based intervention on weight loss: A randomized pilot study. *Obesity, 24*(8), 1653–1659. doi:10.1002/oby.21536

Smith, T. J., Marriott, B. P., Dotson, L., Bathalon, G. P., Funderburk, L., White, A., . . . Young, A. J. (2012). Overweight and obesity in military personnel: Sociodemographic predictors. *Obesity, 20*(7), 1534–1538. doi:10.1038/oby.2012.25

Smith, T. C., Ryan, M. A., Wingard, D. L., Slymen, D. J., Sallis, J. F., Kritz-Silverstein, D., & Millennium Cohort Study Team. (2008). New onset and persistent symptoms of post-traumatic stress disorder self reported after deployment and combat exposures: Prospective population based US military cohort study. *British Medical Journal, 336*(7640), 366–371. doi:10.1136/bmj.39430.638241.AE

Storm, F. A., Heller, B. W., & Mazzà, C. (2015). Step detection and activity recognition accuracy of seven physical activity monitors. *PLoS One, 10*(3), e0118723. doi:10.1371/journal.pone.0118723

Teyhen, D. S., Aldag, M., Edinborough, E., Ghannadian, J. D., Haught, A., Kinn, J., . . . Parramore, D. J. (2014). Leveraging technology: Creating and sustaining changes for health. *Telemedicine Journal and E-Health, 20*(9), 835–849. doi:10.1089/tmj.2013.0328

Toon, E., Davey, M. J., Hollis, S. L., Nixon, G. M., Horne, R. S., & Biggs, S. N. (2016). Comparison of commerical wrist-based and smartphone accelerometers, actigraphy, and

PSG in a clinical cohort of children and adolescents. *Journl of Clinical Sleep Medicine, 12*(3), 343–350. doi:10.5664/jcsm.5580

Troxel, W. M., Germain, A., & Buysse, D. J. (2012). Clinical management of insomnia with brief behavioral treatment (BBTI). *Behavioral Sleep Medicine, 10*(4), 266–279. doi:10.1080/15402002.2011.607200

Valbuena, D., Miltenberger, R., & Solley, E. (2015). Evaluating an Internet-based program and a behavioral coach for increasing physical activity. *Behavior Analysis: Research and Practice, 15*(2), 122–138. doi:10.1037/bar0000013

Wallen, M. P., Gomersall, S. R., Keating, S. E., Wisløff, U., & Coombes, J. S. (2016). Accuracy of heart rate watches: Implications for weight management. *PLoS One, 11*(5), e0154420. doi:10.1371/journal.pone.0154420

Wang, J. B., Cadmus-Bertram, L. A., Natarajan, L., White, M. M., Madanat, H., Nichols, J. F., . . . Pierce, J. P. (2015). Wearable sensor/device (Fitbit One) and SMS text-messaging prompts to increase phsyical activity in overweight and obese adults: A randomized controlled trial. *Telemedicine Journal and E-Health, 21*(10), 782–792. doi:10.1089/tmj.2014.0176

TABLES AND FIGURE

Table 1. Overview of Wearable Activity Trackers Included in the Review

Name	Release Date	Metrics Reported	Placement	Size (cm)	Weight (g)	Cost (US$)
Apple Watch	Apr 2015	Steps, calories, HR	Wrist	AT553: 3.86 (h), 3.33 (w), 1.05 (d) AT554: 4.2 (h), 3.59 (w), 1.05 (d)	40 50	$349.00
Fitbit One	Sep 2012	Steps, distance, calories, active minutes, sleep	Anywhere on body/clip	4.8 (h), 1.9 (w), 1.0 (d)	9	$99.95
Fitbit Flex	May 2013	Steps, distance, calories, active minutes, sleep	Wrist	Small: 14.0–17.6 (c), 1.4 (w) Large: 16.1–20.9 (c), 1.4 (w)	13 15	$99.95
Fitbit Charge	Nov 2014	Steps, distance, calories, active minutes, sleep	Wrist	Small: 14.0–17.0 (c), 2.1 (w) Large: 16.1–20.0 (c), 2.1 (w) Extra large: 19.8–23.0 (c), 2.1 (w)	23	$129.95
Fitbit Charge HR	Jan 2015	Steps, distance, calories, active minutes, sleep, HR	Wrist	Large: 16.1–19.4 (c), 2.1 (w) Extra large: 19.4–23.0 (c), 2.1 (w)	23	$149.95
Garmin vivofit	Mar 2014	Steps, distance, calories, sleep	Wrist	Small: 12.0–17.5 (c), 2.55 (w) Large: 15.2–21.0 (c), 2.55 (w)	25.5	$99.99
Jawbone UP	Nov 2011	Steps, calories, distance, sleep	Anywhere on body/clip	Small: 14.0–15.5 Medium: 15.5–18.0 Large: 18.0–20.0	19 21 23	$59.99
Misfit Shine	Dec 2013	Steps, distance, calories, sleep	Wrist, neck, or body	17.8 (l), 12.7 (w), 2.5 (h)	16	$119.95
Moto 360	Sep 2014	Steps, distance, calories, active minutes, HR	Wrist	4.6 (diameter), 1.15 (h)	49	$249.00
Nike+ FuelBand SE	Nov 2013	Steps, calories	Wrist	Small: 15.1 (c), 1.9 (w) Medium: 17.2 (c), 1.9 (w) Large: 19.7 (c), 1.9 (w)	27–35	$149.95
Qualcomm Toq	Dec 2013	Steps	Wrist	16.2 (l), 15.7 (w), 7.4 (d)	90	$349.99
Samsung Gear S Samsung Gear 2	Oct 2014 Apr 2014	Steps, calories, HR Steps, HR	Wrist Wrist	5.81 (h), 3.99 (w), 1.25 (d) 5.9 (h), 3.8 (w), 0.1 (d)	67 55	$349.00 $249.00

HR, heart rate.

Table 2. Literature Search Criteria by Content Area

Content Area	Area-Specific Search Terms	Exclusion Criteria
Reliability and validity	valid* OR reliab* OR accura* OR comparis* OR comparat*	• Abstracts and/or conference proceedings • Focus on: – discontinued devices – devices not worn on the wrist – devices not commercially available – smartphone applications – research-grade devices – studies unrelated to physical activity – studies that only evaluated sleep – special populations or medical interventions
Sleep	sleep OR sleep monitoring*	• No objective measure of sleep reported • No comparison (gold standard) measure • Focus on: – devices not commercially available – devices that only tracked sleep – clinical populations other than sleep disorders
Behavior change	behavior mod* OR effectiv* OR efficacy OR goal set* OR social learn* OR long term OR sustain* OR maint* OR persist* OR longitudinal	• Focus on: – devices not commercially available – research-grade devices – studies unrelated to physical activity – special populations or medical interventions

Notes: Search date range: January 2009–July 2016. Databases searched included PubMed, Ovid, Scopus, and Google Scholar. All searches included the following terms: fitness track* OR wearable device* OR "wearables" OR physical activity monitor* OR activity monitor* OR activity track* OR wrist worn monitor* OR fitness device*, in addition to area-specific search terms.

Table 3. Characteristics of Studies Included in the Reliability and Validity Review

Lead Author (year)	Sample Size (reliability, validity)	Age (mean±SD, range)	BMI in kg/m^2 (mean±SD, range)	Inclusion Criteria
Bai (2016)	52 (V)	18–65	24.0 17.6–39.9	Apparently healthy adults; no major surgeries in the past year
Case (2015)	14 (V)	28.1 (6.2)	22.7 (1.5)	Apparently healthy adults
Chen (2016)	30 (V)	22.1 (2.2) male 20.1 (1.6) female	21.5 (1.6) male 21.6 (2.3) female	Apparently healthy adults; normal BMI; can ambulate without assistance; regular gait pattern
Diaz (2015)	23 (R, V)	20–54	19.6–29.9	Apparently healthy adults
Diaz (2016)	13 (V)	32 (9.2) 20–54	24.2 19.6–29.9	Apparently healthy adults; can ambulate without assistance; no acute or chronic illness
Dondzila (2016)	19 (V)	24.6 (3.1) 18–30	28.0 (3.8)	No CV, respiratory, musculoskeletal, or metabolic disorders; no limitation to exercise
El-Amrawy (2015)	4 (V)	22–36	Not reported	Apparently healthy adults
Ferguson (2015)	21 (R, V)	32.8 (10.2) 20–59	27.3 (3.2) male 25.5 (5.2) female	Apparently healthy adults
Kooiman (2015)	33 (R, V) in lab 56 (R, V) in field	39 (13.1) male 35 (11.2) female	23.6 (2.2) male 22.5 (2.1) female	Apparently healthy adults
Lee (2014)	60 (V)	24.2 (4.7) female 28.6 (6.4) male	24.3 (2.6), 19.5–28.0 male 21.8 (2.7), 18.1–31.2 female	No major disease; nonsmokers
Nelson (2016)	30 (V)	48.9 (19.4) 18–79	26.3 (5.2) 16.3–38.2	Apparently healthy adults; normal gait pattern; no acute illness or unstable chronic conditions; not pregnant
O'Connell (2016)	15 (R, V)	21.1 (1.1)	23.6 (2.7) male 21.88 (1.81) female	Apparently healthy adults; no CV disease or neurological disorder.
Rosenberger (2016)	40 (V)	36 21–76	Not reported	Apparently healthy adults
Storm (2015)	16 (V)	28.9 (2.7)	23.5 (2.3)	No impairment or morbidity that could interfere with physical activity assessment
Wallen (2016)	22 (R, V)	24.9 (5.6)	Not reported	Apparently healthy adults

BMI, body mass index; CV, cardiovascular; SD, standard deviation.

Table 4. Overview of Validation Studies Included in the Review

Lead Author (year)	Activity	Lab/ Field	Gold Standard Comparison	Metric	Device	Results
Bai (2016)	20 min sedentary; 25 min TM at self-selected speed; 25 min resistance exercise	Lab	Indirect calorimetry using Oxycon Mobile (EE)	kcal/80-min trial	Fitbit Flex	Overestimated EE by 20.4 kcal; *r* = .78; MAPE = 16.8%
					Nike+ FuelBand	Underestimated EE by 42.3 kcal; *r* = .74; MAPE = 17.1%
					Misfit Shine	Overestimated EE by 72.4 kcal; *r* = .71; MAPE = 30.4%
					Polar Loop	Data not included in results; failure of sync features designed to send information from web service and mobile app to device.
Case (2015)	TM at 3.0 mph for 500 and 1,500 steps, each done twice	Lab	Tally counter (steps)	Steps/trial	Fitbit Flex	500 step trial mean = 465.4 (SD 92.1) 1,500 step trial mean = 1378.0 (SD 142.7)
Chen (2016)	4 TM trials at various speeds (54, 80, 107, 134 m•min^{-1}) for 5 min each: 6 simulated activities for 2 hr total: sitting (play computer game, fold laundry), walking (push stroller, carry laptop bag), stair climbing (ascend/descend 3 flights of stairs)	Lab	Step counting from video recording	Steps/trial	Fitbit Flex	MAPE range: 2.5–8.2% (TM trials), 16.7–17.8% (carrying a bag) MAPE <10% (ascending/descending stairs) Overestimated steps while seated folding laundry
					Garmin vivofit	MAPE range: 1.5–4.2% (TM trials) MAPE = 33% (pushing stroller) MAPE <3% (carrying a bag), <10% (ascending/descending stairs) Overestimated steps while seated folding laundry
					Jawbone UP	MAPE range: 2.4–9.6% (TM trials) MAPE = 93.7% (pushing stroller) MAPE <3% (carrying a bag), <10% (ascending/descending stairs) Overestimated steps while seated folding laundry
Diaz (2015)	4 TM trials at various speeds (1.9, 3.0, 4.0, 5.2 mph) for 6 min each	Lab	Step counting from video recording; indirect calorimetry using Ultima CPX (EE)	Steps/min, kcal/min	Fitbit Flex	For steps: *r* = .77–.85, mean difference −26.3 to −2.9 steps For kcal: *r* = .88, mean difference −0.2 to 2.6 kcal
Diaz (2016)	4 TM trials at various speeds (1.9, 3.0, 4.0, 5.2 mph) for 6 min each	Lab	Step counting from video recording, indirect calorimetry using Ultima CPX (EE)	Steps/min, kcal/min	Fitbit Flex	Overestimated step count (2.1 to 15.8% error). Concordance CC .75 (steps); limits of agreement −49.9 to 27.1 steps Underestimated EE (24.5 to 83.4% error). Concordance CC .62 (EE); limits of agreement −0.6 to 4.8 kcal
Dondzila (2016)	4 TM trials at various speeds (80.5, 107.3, 134.1, 160.9 m/min) for 5 min each	Lab	Indirect calorimetry using TrueOne 2400 metabolic cart (EE)	kcal/trial	Fitbit Charge	Overestimated calories at 80.5 m/min Underestimating calories at 107.3 m/min MAPE = 21.4% (80.5 m/min), 11.2% (107.3 m/min), 13.7% (134.1 m/min), 22.5% (160.9 m/min)
El-Amrawy (2015)	Walk 200, 500, and 1,000 steps; repeat 40 times each	Lab	Tally counter (steps), Onyx Vantage 9590 pulse oximeter (HR)	Steps/trial	Misfit Shine	Overall step count accuracy (99.1%) 200 step count accuracy (98.3%), 1,000 step count accuracy (99.7%)
					Samsung Gear 2	Overall step count accuracy (79.8%)
					Jawbone UP	Step count 17.5% precision
					Qualcomm Toq	Overall step count accuracy (97%)
				Steps/trial HR	Apple Watch	200 step count accuracy (99.1%) 1,000 step count accuracy (99.5%) HR accuracy (99.9%), HR precision (5.9%)
				HR	Moto 360	HR accuracy (92.8%)

Lead Author (year)	Activity	Lab/ Field	Gold Standard Comparison	Metric	Device	Results
Ferguson (2015)	48 hr of free-living conditions, no activity restrictions/guidelines	Field	BodyMedia SenseWear model MF (steps, PA, EE); ActiGraph GT3X+ (steps, PA)	Steps/day, MVPA min/day, kcal/day	Jawbone UP	*r* = .97 (steps), .81 (MVPA), .74 (kcal) Mean absolute difference (range) = 806 (steps; −1978 to 2252), 18.0 (MVPA; −4.7 to 96.5), 866 (kcal; −1937 to −94)
					Misfit Shine	*r* = .94 (steps), .79 (MVPA), .79 (kcal) Mean absolute difference (range) = 1,002 (steps; −4,693 to 1,804), 15.2 (MVPA; −79.7 to 36.3), 468 (kcal; −996 to 100)
Kooiman (2015)	TM at 4.8 km/hr for 30 min, twice; 7.5 hr of free-living activities during a working day	Lab/ Field	OptoGait system (steps on TM in lab); activPAL (steps during free-living in field)	Steps/trial	Fitbit Flex	Test-retest reliability on TM via ICC analysis: .81 (good) Mean difference 188 (steps) in lab; MAPE 5.7% Mean difference −150 in free-living; MAPE 3.7%
					Jawbone UP	Test-retest reliability on TM via ICC analysis: .83 (good) Mean difference 34 (steps) in lab; MAPE 1.0% Mean difference −58 in free-living; MAPE 1.4%
					Nike+ FuelBand SE	Test-retest reliability on TM via ICC analysis: .53 (low) Mean difference 598 (steps) in lab; MAPE 18.0% Mean difference 977 in free-living; MAPE 24%
Lee (2014)	13 activities for 69 min total; each activity at 5 min except TM (3 min)	Lab	Oxycon Mobile (EE); ActiGraph GTX3+ worn on hip (EE)	kcal/trial	Jawbone UP	MAPE = 12.2%; *r* = .74 (Oxycon); *r* = .65 (ActiGraph)
					Nike+ FuelBand SE	MAPE = 13.0%; *r* = .40 (ActiGraph)
Nelson (2016)	10 min lying in bed followed by 10 activities (3 sedentary, 4 household, 4 ambulatory/exercise) for 5 min each	Lab	Omron HJ-720IT pedometer (steps); tally counter (steps), indirect calorimetry using COSMED (EE)	Steps/activity EE/activity	Fitbit Flex	For steps: MAPE = 58% (household), 6% (ambulatory), 8% (walking), 8% (jogging) For EE: MAPE = 14% (sedentary), 21% (household), 24% (ambulatory), 53% (walking), 35% (jogging)
O'Connell (2016)	Walk ~2.5 mile route with multiple surface conditions (including stairs) twice	Field	Step counting from video recording, activPAL (steps during walking route)	Steps/trial	Garmin vivofit	MAPE = 5.43% (ceramic tile), 11.33% (ascending stairs), 6.48% (descending stairs) All other MAPE remained within the 5% error zone
Rosenberger (2016)	24 hr of free-living activities, no activity restrictions/guidelines	Field	ActiGraph GT3X+ (time in MVPA), Omron HJ-720IT (steps)	MVPA min/day, steps/day	Jawbone UP	MAPE = 17% (steps), 52% (MVPA) Mean differences of MVPA = 48 min
					Nike+ FuelBand SE	MAPE = 29% (steps), 92% (MVPA) Mean differences of MVPA = 598 min
Storm (2015)	Walk 11 min (indoor, outdoor, and stairs) at self-selected natural, slow, and fast speeds	Lab	Opal sensors placed on each ankle (steps)	Steps/11-min trial	Fitbit One	MAPE = 1.1% (natural), 1.0% (fast) Underestimated steps: −25 (slow), −12 (self-selected), −9 (fast)
					Jawbone UP	MAPE = 10.1% (slow), 2.5% (natural), 2.1% (fast) Underestimated steps: −35 (slow), −12 (natural), −9 (fast)
Wallen (2016)	1 hr protocol (5 min supine, 5 min seated, 5 min standing, 9 min TM, 5 min seated, 18 min cycle, 5 min seated) with 1 min rest between activities	Lab	ECG (HR), indirect calorimetry using METAMAX 3B (EE); step counting from video recording	HR kcal/trial Steps/trial	Apple Watch	Underestimated steps, EE, HR *r* = .70 (steps), .16 (EE), .95 (HR) Mean difference: −123.1 ± 55.6 (EE), −1.3 ± 4.4 (HR)
					Fitbit Charge HR	Underestimated steps, EE, HR *r* = .67 (steps), .64 (EE), .81 (HR) Mean difference: −224.6 ± 59.1 (EE), −9.3 ± 8.5 (HR)
					Samsung Gear S	Underestimated steps, EE, HR *r* = .88 (steps), .86 (EE), .67 (HR) Mean difference: −26.1 ± 24.2 (EE), −7.1 ± 10.3 (HR)

CC, correlation coefficient; ECG, electrocardiogram; EE, energy expenditure; HR, heart rate; ICC, intraclass coefficient; MAPE, mean absolute percentage error; MVPA, moderate to vigorous physical activity; PA, physical activity; SD, standard deviation; TM, treadmill.

Table 5. Overview of Sleep Studies Included in the Review

Lead Author (year)	Devices	Gold Standard	Subjects	Study Site	Measure(s) Tested	Sensitivity	Specificity	Results
de Zambotti (2015)[a]	Jawbone UP	PSG	*N* = 28, middle-aged women only, mean = 50.1 y, 12 with insomnia, 4 with PLMS	Lab	TST, SOL, WASO	.96	.37	High sensitivity, low specificity; overestimated TST and SOL; underestimated WASO
de Zambotti (2015)[b]	Jawbone UP	PSG	*N* = 65, healthy adolescents, mean = 18.8 y, range = 12–22 y	Lab	TST, SOL, WASO, SE	N/A	N/A	Overestimated TST and SE; underestimated WASO; no difference in SOL
de Zambotti (2016)	Fitbit Charge HR	PSG	*N* = 32, healthy adolescents, mean = 17.3 y, range = 12–21 y	Lab	TST, SOL, WASO, SE	.97	.42	Significantly overestimated TST and SE; underestimated WASO
Mantua (2016)	Basis Health Tracker; Misfit Shine; Fitbit Flex; Withings Pulse O2	PSG, actigraphy	*N* = 40, healthy young adults, mean = 22.4 y, range = 18–30 y	Home	TST, SE, light, deep	N/A	N/A	TST from all devices correlated with PSG, noncorrelated with SE; light and deep sleep results were mixed; only Withings Pulse O2 correlated with light and deep sleep but was <40%
Meltzer (2015)	Fitbit Ultra	PSG, actigraphy	*N* = 63, children/adolescents, mean = 9.7 y, range = 3–17 y, 23% had mild OSA, 16% had moderate/severe OSA	Sleep clinic	TST, SE	Normal mode: .86 Sensitive mode: .70	Normal mode: .52 Sensitive mode: .79	Tradeoffs in sensitivity, specificity, and over/underestimated TST, SE, and WASO between modes; neither mode showed optimal performance
Montgomery-Downs (2012)	Fitbit (classic)	PSG, actigraphy	*N* = 24, healthy young adults, mean = 26.1 y, range = 19–41 y	Lab	TST, SE	.98	.20	High sensitivity, low specificity; significantly overestimated TST and SE
Rosenberger (2016)	Fitbit One; Jawbone UP	General Sleep Zmachine	*N* = 40, adults, mean = 36 y, range = 21–76 y	Home	TST	N/A	N/A	Both overestimated TST
Toon (2016)	Jawbone UP	PSG, actigraphy	*N* = 78, children/adolescents, mean = 8.4 y, range = 3–18 y, 41% with mild OSA, 28% with moderate/severe OSA, 6% with PLMS	Sleep clinic	TST, SOL, WASO, SE	.92	.66	Equivalent sensitivity, specificity, and accuracy without PSG; device outperformed actigraphy for SOL

OSA, obstructive sleep apnea; PLMS, periodic limb movement disorder; PSG, polysomnography; SE, sleep efficiency; SOL, sleep onset latency; TST, total sleep time; WASO, wake after sleep onset.

[a] de Zambotti, Claudatos, Inkelis, Colrain, & Baker (2015)

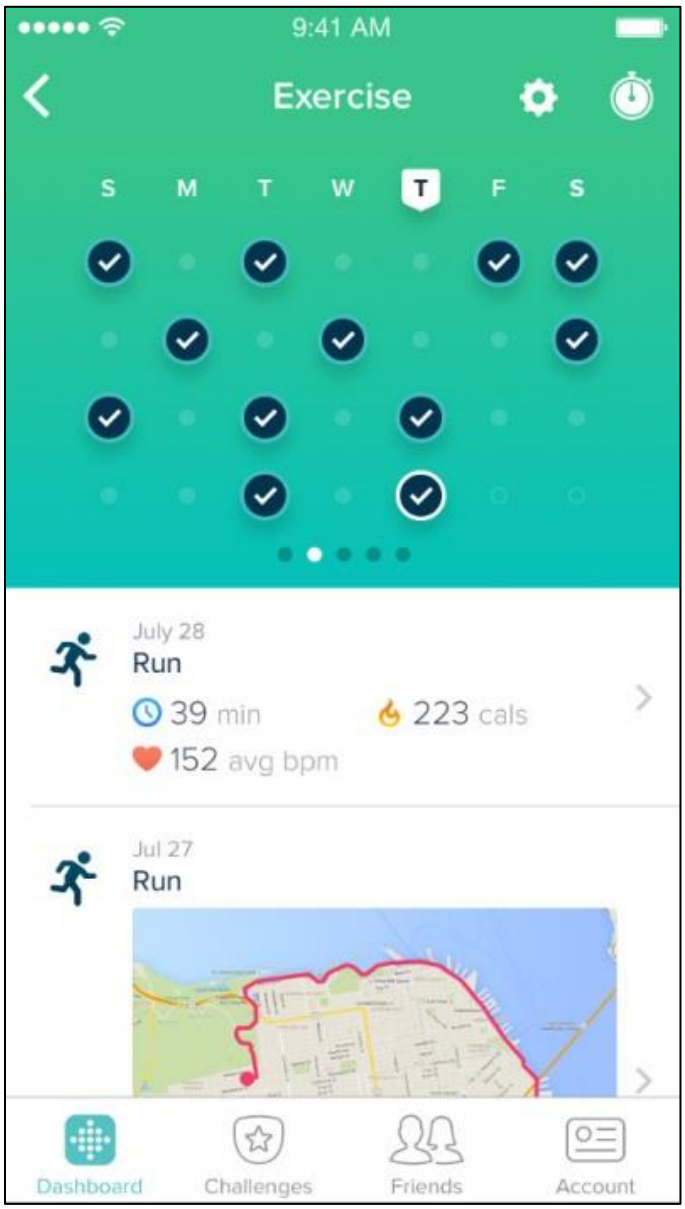
[b] de Zambotti, Baker, & Colrain (2015)

Table 6. Overview of Behavior Change Studies Included in the Review

Lead Author (year)	Subjects	Purpose	Devices	Design	Results
U.S. Army (2015–ongoing)*	*N* = 2,200, U.S. Army soldiers (2 brigades)	To enhance soldier and unit performance through improvements in sleep, PA, and nutrition, with the overall goal to improve Force readiness and increase resilience	Fitbit Charge HR	26 weeks, 3 stages: 1. P3 program alone 2. WAT + P3 program 3. Control (no P3 program, no WAT)	WAT usage fell to about 30% after 2 months (8 weeks) into the study. Only 29–42% of participants in the WAT + P3 group met all activity goals.
Finkelstein (2016)	*N* = 800, desk-bound office workers; 53–61% of subjects were overweight or obese. Age = 21–65 y	To test whether WAT, alone or in combination with cash incentives or charitable donations, could increase PA and improve health outcomes.	Fitbit Zip	6 months, 4 conditions: 1. WAT only 2. WAT + cash incentive 3. WAT + charity incentive 4. Control	Cash incentive was most effective at increasing MVPA at 6 months; effect was not sustained after incentives were discontinued. No improvement in health outcomes with any group.
Groendal (2014)	*N* = 33, inactive adults who did not meet PA recommendations Mean age = 45±12 y BMI = 29.3±6.8 kg/m^2	To study changes in inactive adults' level of PA after a WAT intervention compared with active controls	Fitbit	8 weeks + 6-month follow-up: 1. WAT (intervention group) 2. Control (no WAT)	Wearing the WAT for 8 weeks increased PA in inactive individuals to the point of meeting/exceeding ACSM PA recommendations. Intervention group self-reported PA remained elevated 6 months post-intervention despite not wearing the WAT since week 8.
Jakicic (2016)	*N* = 471, overweight or obese young adults Age = 18–35 y BMI = 25–40 kg/m^2	To determine if a technology-enhanced weight loss intervention would result in greater weight loss than that found with a standard behavioral weight loss intervention	Fitcore; BodyMedia	24 months total (6-month intervals), 2 groups: 1. Standard use: self-monitoring diet & PA using website 2. WAT-enhanced: use WAT and web interface to monitor diet/PA	Standard intervention group had lost more weight in 24 months than the WAT-enhanced intervention group. Both groups had improvements in body composition, fitness, PA, and diet, but there were no significant differences between groups.
Ross (2016)	*N* = 80, overweight or obese adults Age = 18–70 BMI = 27–40 kg/m^2	To examine the impact of WATs, provided with and without brief phone-based interventions, on weight loss in overweight and obese adults, compared with standard pedometers	Fitbit Zip (waist worn)	6 months, no follow-up, 3 groups: 1. Pedometer only 2. WAT only 3. WAT + phone-based counseling	WAT + phone-based counseling group lost more weight over 6-month intervention than either pedometer only or WAT-only groups.
Valbuena (2015)	*N* = 7, overweight and obese adults BMI = 28–40.7 kg/m^2 Age = 44–57 y	To evaluate the effectiveness of a WAT program, with and without a behavioral coach, in a multiple-baseline across-participant design	Fitbit One	3 phases of various lengths (147–449 days total) 1. Baseline: no access to WAT data 2. Access to WAT data 3. Access to WAT data + individualized behavioral coaching	WAT program alone increased PA for 3 of 7 participants; addition of the behavioral coach further increased mean step count for all 6 participants.
Wang (2015)	*N* = 67, overweight and obese adults BMI ≥25 kg/m^2 Age = 18–69 y	To test the effects on PA level of a WAT-based intervention that delivered simple prompts using text messaging in conjunction with the WAT for self-monitoring	Fitbit One	6 weeks, 2 groups: 1. WAT 2. WAT + 3 daily SMS-based PA prompts	WAT-only group significantly increased PA compared with WAT + PA prompts. The positive effect on PA of the text messaging only lasted during the initial first week.

ACSM, American College of Sports Medicine; BMI, body mass index; MVPA, moderate to vigorous physical activity; P3, U.S. Army Performance Triad; PA, physical activity; WAT, wearable actvity tracker.

* Source: T. Brahmbhatt (personal communication, February 18, 2016)



https://www.fitbit.com/app

Figure 1. Screenshot from a Global Positioning System-Enabled Wearable Activity Tracker Application.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
14-03-2017	Technical	January 2009- July 2016

4. TITLE AND SUBTITLE

Wearable Activity Tracker Literature Review (January 2009 – July 2016)

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Shawn E. Soutiere, Brennan D. Cox, Melissa D. Laird, Rachel R. Markwald, Jay H. Heaney, Evan D. Chinoy, Rita G. Simmons

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Commanding Officer
Naval Health Research Center
140 Sylvester Rd
San Diego, CA 92106-3521

8. PERFORMING ORGANIZATION REPORT NUMBER

17-38

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Commanding Officer
Naval Medical Research Center
503 Robert Grant Ave
Silver Spring, MD 20910-7500

Chief, Bureau of Medicine and Surgery
(MED 00), Navy Dept
7700 Arlington Blvd Ste 5113
Falls Church, VA 22042-5113

10. SPONSOR/MONITOR'S ACRONYM(S)

BUMED/NMRC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

17-311

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Wearable activity trackers (WATs) have become increasingly popular for their purported ability to provide users with real-time, tailored information about their daily health-related activities. These devices are portrayed as user-friendly and beneficial to well-being, which raises the question of their potential application for health promotion among military populations. However, the rapid evolution of WAT technology is outpacing efforts to test, evaluate, and validate these devices. Therefore, the purpose of this systematic review was to summarize the evidence on WAT reliability and validity for measuring physical activity (PA) and sleep, and the utility of incorporating WATs in behavior modification programs.

15. SUBJECT TERMS

WAT, health, physical activity, measure, information

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Commanding Officer
U	U	U	UU	33	19b. TELEPHONE NUMBER *(Include area code)* COMM/DSN: (619) 553-8429