Commercial Devices Provide Estimates of Energy Balance with Varying Degrees of Validity in Free-Living Adults

Robin P Shook,[1,2] Hung-Wen Yeh,[3] Gregory J Welk,[4] Ann M Davis,[5] and Daniel Ries[6]

[1]Department of Pediatrics, Center for Children's Healthy Lifestyles and Nutrition, Children's Mercy Kansas City, Kansas City, MO, USA; [2]School of Medicine, University of Missouri-Kansas City, Kansas City, MO, USA; [3]Department of Pediatrics, Health Services and Outcomes Research, Children's Mercy Kansas City, Kansas City, MO, USA; [4]Department of Kinesiology, Iowa State University, Ames, IA, USA; [5]Department of Pediatrics, Center for Children's Healthy Lifestyles and Nutrition, University of Kansas Medical Center, Kansas City, KS, USA; and [6]Sandia National Laboratories, Albuquerque, NM, USA

ABSTRACT

Background: The challenges of accurate estimation of energy intake (EI) are well-documented, with self-reported values 12%–20% below expected values. New approaches rely on gold-standard assessments of the other components of energy balance, energy expenditure (EE) and energy storage (ES), to estimate EI.

Objectives: The purpose of this study was to evaluate the validity, repeatability, and measurement error of consumer devices when estimating energy balance in a free-living population.

Methods: Twenty-four healthy adults (14 women, 10 men; mean ± SD age: 30.7 ± 8.2 y) completed two 14-d assessment periods, including assessments of EE and ES using gold-standard [doubly labeled water (DLW) and DXA] and commercial devices [Fitbit Alta HR activity monitor (Alta) and Fitbit Aria wireless body composition scale (Aria)], and of EI by dietician-administered recalls. Accuracy and validity were assessed using Spearman correlation, interclass correlation, mean absolute percentage error, and equivalency testing. We also applied linear measurement error modeling including error in gold-standard devices and within-subject repeated-measures design to calibrate consumer devices and quantify error.

Results: There was moderate to strong agreement for EE between the Fitbit Alta and DLW at each time point ($r_S = 0.82$ and 0.66 for Times 1 and 2, respectively). There was weak agreement for ES between the Fitbit Aria and DXA ($r_S = 0.15$ and 0.49 for Times 1 and 2, respectively). Correlations between methods to assess EI ranged from weak to strong, with agreement between the DXA/DLW-calculated EI and dietary recalls being the highest ($r_S = 0.63$ for Time 1 and 0.73 for Time 2). Only EE from the Fitbit Alta at Time 1 was equivalent to the DLW value using equivalency testing.

Conclusions: Commercial devices provide estimates of energy balance in free-living adults with varying degrees of validity compared to gold-standard techniques. EE estimates were the most robust overall, whereas ES estimates were generally poor. *J Nutr* 2021;00:1–9.

Keywords: energy balance, consumer devices, measurement error modeling, energy intake, energy expenditure, energy storage

Introduction

The challenges of accurate estimation of energy intake (EI) are well-documented (1–3). Recently, mathematical models have been formulated based on the principles of the first law of thermodynamics [rate of energy storage (ES) = rate of EI − rate of energy expenditure (EE)] (4) that allow researchers to estimate energy balance. For example, if one is able to accurately measure 2 of the variables of the energy balance equation (e.g., changed ES and rate of EE), it is mathematically possible to solve for the third (e.g., EI). Based on a variety of existing data sets containing EE, EI, and changes in ES [e.g., body composition using a 2-compartment model of fat mass (FM) and fat-free mass (FFM)] during periods of overfeeding (5) or caloric restriction (6), researchers have developed and refined a technique termed the intake-balance method to estimate EI (7–9). The result is a simple, easy-to-use equation that offers great promise in the quest for estimating EI using objectively measured methods.

A current limitation in using the intake-balance method to estimate EI is the feasibility in measuring EE and ES. Although both can be accomplished with a high degree of accuracy, the gold-standard method of assessment [doubly labeled water (DLW) for EE, DXA for ES] of each is too costly and resource-intensive for most applications. Consumer devices for assessing physical activity and body composition are

Manuscript received March 17, 2021. Initial review completed May 11, 2021. Revision accepted August 26, 2021.
First published online 0, 2021; doi: https://doi.org/10.1093/jn/nxab317.

affordable, easy to use, and popular (an estimated 45 million were sold in 2017) (10) but have varying levels of validity and reliability (11–13). Although market turnover of the devices generally outpaces scientific validity and reliability output, it is generally accepted that commercial devices perform adequately for outcomes such as steps per day and are less valid for minutes of physical activity (14). There is considerable public interest in body weight management so it is important to evaluate the validity of computed estimates of energy balance based on this methodology using consumer-based devices.

The purpose of this study was to explore the validity of consumer devices to estimate the 3 components of the energy balance model (EE, ES, and EI) under free-living conditions. Our research group has successfully developed a system of assessment methodologies and statistical techniques that can accurately assess energy balance and estimate EI using research-grade but non-gold-standard techniques (15, 16). We sought to expand these techniques from use with research-grade devices to consumer devices.

Methods

Design

Twenty-four participants (14 women) were healthy adults aged >21 y (range: 21–52 y; BMI range: 19.1–34.8 kg/m^2) (**Supplemental Figure 1**). All participants had to have an in-home Wi-Fi network and access to a smartphone that could operate mobile applications ("apps"). Sample size was determined using previous estimates by the study team (16) and review of the existing literature. The sample size of $n = 24$ participants would provide estimates with a 2-sided margin of error of $t_{0.975,\ \mathrm{df}\,=\,\mathrm{N}-1}/\sqrt{N} = 0.42$ SD at a 95% confidence level. All study procedures were approved by the Children's Mercy Institutional Review Board and all participants provided consent before data collection. All participants completed the study protocol between 1 January, 2018 and 31 July, 2018. The protocol included 2 identical 14-d assessment periods, separated by a 14-d washout period (**Figure 1**), and included assessments of EE and ES using both gold-standard and commercial devices. Participants arrived fasted on the morning of day 0 before 09:00, and immediately provided a urine sample which served as the background sample for the DLW procedure. They then changed into hospital scrubs and socks, were measured on a certified scale and stadiometer for height and weight, and completed a DXA scan (Lunar iDXA, GE Healthcare) to assess body composition. Immediately afterwards they were assessed for weight and body composition using the Fitbit Aria wireless body composition scale (Fitbit, Inc.; described below), then DLW dosing occurred. Fourteen days later, participants assessed their body weight and composition using the Fitbit Aria in their home, then returned to the laboratory before 09:00 and repeated the DXA scan. During the period in between, participants self-assessed their body weight and composition each morning, upon waking and postvoid, using the Fitbit Aria smart scale. They also continuously wore a Fitbit Alta HR™ activity monitor (Aria; Fitbit, Inc.; described below) at all times, including sleeping, except while showering, bathing, or swimming. Self-reported dietary intake (described below) was assessed using 3 dietician-administered recalls occurring over the telephone. This 14-d protocol was repeated in its entirety after a 14-d washout period.

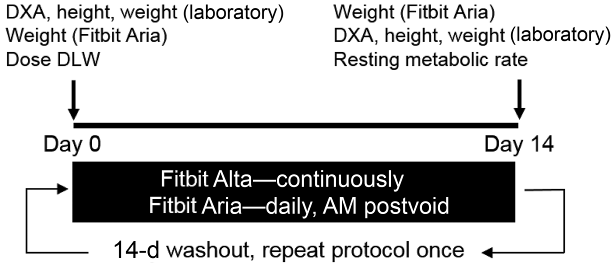
DXA, height, weight (laboratory)
Weight (Fitbit Aria)
Dose DLW
Weight (Fitbit Aria)
DXA, height, weight (laboratory)
Resting metabolic rate
Day 0
Day 14
Fitbit Alta—continuously
Fitbit Aria—daily, AM postvoid
14-d washout, repeat protocol once

FIGURE 1 Study design. Alta, Alta HR activity monitor; Aria, Aria wireless body composition scale; DLW, doubly labeled water.

Measures

The Fitbit Aria is an electric scale which uses bioelectrical impedance to assess body composition and wirelessly uploads the data to a user's Fitbit account. The Fitbit Alta uses a 3-axis accelerometer to track movement, an altimeter to track change in altitude, and light-emitting diodes (LEDs) to estimate heart rate. This information, along with user characteristics such as body weight, sex, and age, are input into proprietary algorithms to estimate steps; minutes of activity occurring at sedentary, light, moderate, and very active intensities; and EE, and are uploaded to a user's Fitbit account. The study team created user accounts for all participants using a generic email address without using their first or last name. All data from both the scale and activity monitor were then aggregated using a comprehensive data management platform (Fitabase™, Small Steps Lab LLC) and downloaded at study completion.

The DLW method was the gold-standard technique used to assess EE. The dose of DLW administered was 1 g/kg of body weight, which was consumed from a sterile bottle and followed by drinking 100 mL of tap water from the same bottle to ensure all of the isotopes were consumed. On day 0, participants provided a background urine sample upon arriving in the laboratory, 2–3 samples after the DLW dosing which were discarded, and a final sample 4.5–5.0 h after dosing. On day 7, participants collected a sample at home in a collection cup and stored it in the refrigerator until they returned to the laboratory. On day 14, the participant arrived in the laboratory and provided a final urine sample. All samples were logged for date and time of day collected, separated into aliquots in 10-mL tubes, and stored in a −80°C freezer. All samples were sent to the Mass Spectrometry Core at Pennington Biomedical Research Center at study completion to determine EE via mass spectroscopic analysis of urine specimens for deuterium and oxygen-18 (17, 18).

EI (hereafter "reported EI") was measured concurrently with EE over a 14-d period using interviewer-administered 24-h dietary recall interviews (24HRs) utilizing the Nutrient Data System for Research software (NDSR®, version 2017) (19). Three 24HRs were administered on randomly selected, nonconsecutive days (including ≥ 1 weekend day) to minimize preparation that could bias recall by the participants (20). The dietary recalls were collected by a team of experienced registered dietitians using a multipass approach, which utilizes prompting to reduce food omissions, and standardizes the interview methodology across interviewers (21). Before data collection, study participants underwent a brief training session (10–15 min) on how to estimate

Supported in part by an unrestricted research grant from the International Life Sciences Institute of North America (ILSI NA) (to RPS). ILSI NA had no role in any aspect of the study, study design, or manuscript development. ILSI NA is a public, nonprofit foundation that provides a forum to advance understanding of scientific issues related to the nutritional quality and safety of the food supply by sponsoring research programs, educational seminars and workshops, and publications. ILSI NA receives support primarily from its industry membership. The opinions expressed herein are those of the authors and do not necessarily represent the views of the funding organization.

Author disclosures: The authors report no conflicts of interest.

Supplemental Figure 1, Supplemental Method 1, and Supplemental Table 1 are available from the "Supplementary data" link in the online posting of the article and from the same link in the online table of contents at https://academic.oup.com/jn/.

Address correspondence to RPS (e-mail: rpshook@cmh.edu).

Abbreviations used: Alta, Alta HR activity monitor; Aria, Aria wireless body composition scale; DLW, doubly labeled water; EE, energy expenditure; EI, energy intake; ES, energy storage; FFM, fat-free mass; FM, fat mass; ICC, intraclass correlation coefficient; LMEM, linear measurement error modeling; MAPE, mean absolute percentage error; RMSE, root mean square error; TOST, two-one-sided-test; 24HR, 24-h dietary recall interview.

portion sizes of commonly eaten foods, and were provided with a 2-dimensional, validated food portion visual to assist in identifying the food amounts consumed (22). Owing to data recording error, dietary recall data were not recorded for 3 participants during Time 1 and 5 participants during Time 2. In total, dietary recalls were not recorded for 3 participants at either time point, and 2 participants only recorded data at Time 2. These participants were excluded from all dietary analyses. Of those with reported dietary data, 100% completed ≥2 of 3 possible recalls at Time 1 (19 of 21 completed all 3) and Time 2 (18 of 19 completed all 3).

EI (hereafter "calculated EI") also was estimated over each 14-d period based on the energy balance equation, represented as:

$$\mathrm{ES} = \mathrm{EI} - \mathrm{EE} \tag{1}$$

where ES represents changed body energy stores; EI represents the rate of EI; and EE represents the rate of EE. When 2 terms of the energy balance equation are known, it is possible to solve for the third term. Thus, estimated EI was calculated based on the following validated equation (6, 15, 23, 24):

$$\text{calculated EI} = 1020\frac{\Delta\mathrm{FFM}}{\Delta t} + 9500\frac{\Delta\mathrm{FM}}{\Delta t} + \mathrm{EE} \tag{2}$$

where ΔFFM and ΔFM represent the change in each variable as measured by DXA between day 1 and day 14; Δt represents days between the start and end of the assessment period; 1020 represents the energy density in kilocalories of FFM per kilogram and 9500 represents the energy density of FM per kilogram, both based on established values (25); and EE represents daily EE as measured by either DLW or the Fitbit Alta during the assessment period.

Statistical analyses

Demographic characteristics were summarized by means ± SDs for all participants and by participants' gender. EI, EE, and ES measures obtained from gold-standard and commercial devices were also summarized by means ± SDs for each time point; body weight and composition were summarized by time point and by day (days 0 and 14).

Consistency between devices and recalls was evaluated by Spearman correlation coefficient (r_s) and agreement was assessed by absolute agreement intraclass correlation coefficient (ICC), mean absolute percentage error (MAPE), and the agreement limits (26) at each time point. Considering the small sample size, we chose Spearman instead of Pearson to alleviate sensitivity due to outliers, and applied the small-sample correction in computing agreement limits (27).

Equivalence between mean energy measures obtained from devices and self-report (EI only) was evaluated by the two-one-sided-test (TOST) to formally test the equivalence in group-level agreement (28–30). In the TOST, an equivalence zone with a lower and an upper bound [defined as 10% below and above, i.e., 90%–110% of, the mean value from the gold-standard device here following Calabro et al. (28)] is used to form 2 composite null hypotheses: the true difference is less than the lower bound or is above the upper bound. When both null hypotheses are rejected, one can conclude that the true difference falls within the equivalence zone, i.e., it is small enough to be claimed as equivalent. For visualization purposes, we computed 90% (instead of the conventional 95% because of the two-one-sided null hypotheses) CIs for measures from the commercial device and against the equivalence zone of the gold-standard device; in this way, the CI should completely fall within the equivalence zone when the TOST is significant at the 5% level. As noted by Dixon et al. (29), this differs from traditional hypothesis testing in that it reverses the traditional null hypothesis to specify that 2 methods are not the same. For the null hypothesis to be rejected (i.e., 2 methods are not equivalent) in a 95% equivalence test, 90% CIs of gold-standard or 7 non-gold-standard assessments must be completely included within a prespecified zone of equivalence [±10% of the mean in this study, as previously used by Calabro et al. (28)]. This corresponds to rejecting 2 one-sided tests: a lower end of a 95% CI greater than a lower boundary (i.e., −10%) of the equivalence zone and an upper end of a 95% CI smaller than an upper boundary (i.e., +10%) of the equivalence zone. Considering that failing to reject the null hypothesis in TOST does not imply the measures were different, we also compared measures between devices and recall by paired t test with Hedges' g as the effect size measure.

The aforementioned analyses used measurements from gold-standard devices as comparative values and were conducted for each measure and at each time point separately. We also integrated all measurements at both time points and applied linear measurement error modeling (LMEM) (16, 31) which allowed measurement errors in gold-standard devices and the full usage of information from the within-subject repeated-measures design. In LMEM, we assumed that *1*) true EE and true ES were latent variables following a bivariate Gaussian distribution; *2*) measurements from gold-standard devices followed Gaussian distributions with mean equal to the true EE and ES values and with nonzero variance, i.e., gold-standard devices provided unbiased measurements with measurement error; *3*) measurements from commercial devices followed Gaussian distributions with mean as a function of true EE and ES, age, sex, and BMI, and with nonzero variance; and *4*) EI, expenditure amount, and metabolism rate (ES) among the participants remained constant during the data collection period. Estimation of LMEM parameters was conducted by Bayesian inference (see **Supplemental Method 1** for details and choices of prior distributions and hyperparameter values). The resulting posterior medians of true EE and ES were then used to calibrate gold-standard and commercial devices and to quantify corresponding measurement errors summarized by root mean square error (RMSE; each observation was first subtracted from its corresponding posterior median of true value, then we squared the difference, and then averaged the squared differences across all participants and time points).

The analyses were conducted on the statistical software R (R Core team 2019) version 3.5.2 using the packages irr (32) for ICC at each time point, equivalence (33) for TOST, and rstan for LMEM (34). Bland–Altman plots and analysis were conducted in GraphPad Prism version 9.0.0.

Results

Table 1 presents participant characteristics. Table 2 presents energy balance descriptive data. The Fitbit Aria assessed body weight as higher than did the laboratory scale and underestimated FM compared with DXA at each time point, although these differences were only significant at $P < 0.05$ on Time 1, day 0, (body weight) and on Time 1, day 14 and Time 2, days 0 and 14 (FM). FM differences between DXA and the Fitbit Aria were relatively small for all visits (<0.4 kg) except day 0 of the first assessment period (1.2 kg); the same pattern existed for FFM.

Table 3 and Figures 2–4 summarize agreement and consistency between devices and dietary recall. The Fitbit Alta demonstrated moderate to strong agreement and correlation as compared to DLW (agreement ICC = 0.81 and 0.64, $r_s = 0.82$ and 0.66 for Times 1 and 2, respectively) (Table 3) with small degrees of overestimation (mean bias 17 and 76 kcal/d, MAPE = 9.3% and 12.5%). There was weak agreement for ES between the Fitbit Aria and DXA at each time point ($r_s = 0.15$ and 0.49 for Times 1 and 2, respectively). Correlations between various methods to assess EI ranged from weak to strong, with consistency between the DXA/DLW-calculated EI and dietary recalls being the highest ($r_s = 0.63$ for Time 1 and 0.73 for Time 2). Agreements between EI estimates from the Fitbit devices and either DLW/DXA or dietary recalls were quite variable by each time point. The correlation between Fitbit devices and DLW/DXA was 0.27 at Time 1, but improved to 0.53 at Time 2. The opposite was true for Fitbit devices and dietary recalls, with moderate agreement at Time 1 ($r_s = 0.63$) but weak agreement at Time 2 ($r_s = 0.42$). Bland–Altman plots are presented in

TABLE 1 Participant characteristics overall and by sex[1]

Characteristics	All ($n = 24$)	Women ($n = 14$)	Men ($n = 10$)
Age, y	30.7 ± 8.2	29.5 ± 6.2	32.4 ± 10.6
Height, cm	171.8 ± 7.8	168.3 ± 7.8	176.7 ± 4.5
Body weight, kg	73.1 ± 11.6	68.4 ± 10.0	79.6 ± 10.6
BMI, kg/m^2	24.8 ± 4.0	24.3 ± 4.6	25.5 ± 3.0
Body fat,[2] %	29.0 ± 10.0	33.7 ± 8.8	22.5 ± 7.8

[1] Values are means ± SDs.
[2] DXA day 0.

Figure 2 for EE and ES and in Figure 3 for EI. Bias (the mean of the differences) varied considerably based on device and time point. For example, bias was low for EE at Times 1 and 2 between DLW and the Fitbit Alta (−17.1 and −76.1 kcal/d, respectively), whereas it was quite high at Time 1 for ES but not at Time 2 (337.6 and −16.8 kcal/d, respectively). For EI (Figure 3), the lowest bias was between the Fitbit devices and dietary recall at Time 1 (−19.7 kcal/d) although a clear inverse relation existed, whereas the highest bias was between the gold-standard and Fitbit devices at Time 1 (320.0 kcal/d) although this relation was much lower at Time 2 (−92.8 kcal/d).

Table 4 presents interclass correlations between the gold-standard and commercial devices. Similar associations were observed at each time point as described with Spearman correlations. In addition, when both time points were combined, data suggested the agreement between devices was excellent for EE (0.76), weak for ES (0.27), and fair for calculated EI (0.44) according to Cicchetti (35).

Results of testing for difference and equivalence are presented in Table 4 and Figure 4 (complete data and values are available in **Supplemental Table 1**). Hedges' g (in absolute value) ranged from 0.02 to 0.38 (Table 4). The only values that appeared equivalent were EE derived from DLW and the Fitbit Alta: the TOST suggested their "equivalence" ($P = 0.0005$ and 0.03 at Times 1 and 2, respectively) and the 2-sided paired t test provided consistent results and did not reject the null hypothesis of no difference ($P = 0.79$ and 0.40, respectively). For other comparisons between devices and with dietary recalls, both paired t test and TOST failed to reject their null hypotheses, leaving the results inconclusive. Among them, Hedges' g appeared small for EI between Fitbit and recall at Time 1 ($g = 0.02$) and for ES between the Fitbit Aria and DXA at Time 2 ($g = 0.03$), suggesting that the differences might be small and a significant TOST might have been achievable if we had had larger sample sizes.

Finally, **Figure 5** presents RMSEs and 95% CIs derived from LMEM for each assessment technique including both time points. If our assumptions hold, RMSE suggested that the gold-standard devices were not free of measurement error, with greater measurement error in DLW (RMSE = 524 kcal/d) than DXA (406 kcal/d). The Fitbit Alta demonstrated a similar level of measurement error as DLW; however, the Fitbit Aria appeared to have substantially greater measurement error (875 kcal/d). EI measures inferred from gold-standard and commercial devices had greater measurement errors (732 and 1026 kcal/d, respectively) than their corresponding EE and ES measures, which was anticipated because intake was estimated by the sum of the other 2. Interestingly, EI obtained from dietary recalls appeared to be more precise than either gold-standard or commercial devices (624 kcal/d).

Discussion

This study evaluated the utility of the "intake-balance" methodology using widely used consumer-based devices (Fitbit

TABLE 2 Energy balance descriptive information of healthy adult participants collected during two 14-d assessment periods, Time 1 and Time 2[1]

	Time 1		Time 2	
	Day 0	Day 14	Day 0	Day 14
Body weight,[2] kg	73.1 ± 11.6	72.9 ± 11.7	73.1 ± 11.5	72.9 ± 11.5
Body weight,[3] kg	73.6 ± 11.8*	73.1 ± 12.0	73.3 ± 11.2	73.1 ± 11.4
FM,[4] kg	21.4 ± 8.7	21.2 ± 8.4	21.1 ± 8.3	21.0 ± 8.4
FM,[3] kg	20.2 ± 7.8	19.8 ± 7.5**	19.8 ± 7.5**	19.8 ± 7.5
FFM,[4] kg	52.5 ± 10.4	52.5 ± 10.7	52.8 ± 10.2	52.6 ± 10.5
FFM,[3] kg	53.4 ± 9.4	53.4 ± 10.2	53.5 ± 9.7	53.3 ± 9.7
EE,[5] kcal/d	—	2584 ± 516	—	2490 ± 490
EE,[6] kcal/d	—	2601 ± 490	—	2567 ± 520
ΔES,[4] kcal/d	—	− 145 ± 434	—	− 89 ± 413
ΔES,[3] kcal/d	—	− 483 ± 1102	—	− 72 ± 571
Calc EI,[7] kcal/d	—	2439 ± 798	—	2402 ± 677
Calc EI,[8] kcal/d	—	2119 ± 1204	—	2494 ± 839
EI, kcal/d[9]	—	2117 ± 538	—	2268 ± 625

[1] Values are means ± SDs. *$P < 0.05$ compared with Laboratory at the same time point; **$P < 0.05$ compared with DXA at the same time point. ΔES refers to change over each 14-d assessment period. Alta, Alta HR activity monitor; Aria, Aria wireless body composition scale; Calc, calculated; DLW, doubly labeled water; EE, energy expenditure; EI, energy intake; ES, energy storage; FFM, fat-free mass; FM, fat mass.
[2] Laboratory.
[3] Fitbit Aria.
[4] DXA.
[5] DLW.
[6] Fitbit Alta.
[7] DXA/DLW.
[8] Fitbit Aria/Alta.
[9] Self-report.

TABLE 3 Agreement and consistency between devices and dietary recall[1]

Measure	Contrast	Time	ρ[2]	ICC[3]	Mean bias[4]	MAPE[4]	95% Agreement limits, kcal/d	
Expenditure	Fitbit Alta vs. DLW	1	0.82	0.81	17	9	− 536	571
		2	0.66	0.64	76	13	− 677	829
Storage	Aria vs. DXA	1	0.15	0.11	− 338	2087	− 2288	1613
		2	0.49	0.39	17	1830	− 956	990
Intake	Fitbit vs. DLW/DXA	1	0.27	0.26	− 320	44	− 2492	1852
		2	0.53	0.50	93	27	− 1247	1433
	DLW/DXA vs. Recalls	1	0.63	0.68	220	21	− 743	1182
		2	0.74	0.65	102	17	− 872	1075
	Fitbit vs. Recalls	1	0.63	0.48	20	44	− 1810	1849
		2	0.42	0.49	213	29	− 1021	1446

[1] Alta, Alta HR activity monitor; Aria, Aria wireless body composition scale; DLW, doubly labeled water.

[2] Spearman correlation coefficient.

[3] Absolute agreement intraclass correlation coefficient.

[4] Mean absolute percentage error.

Alta and Fitbit Aria) to capture the key components. The primary finding from this study is that commercial devices have differential validity for capturing the 3 components of the energy balance model (EE, ES, and EI). EE estimates were the most robust overall, whereas ES estimates were generally poor. As a result of this variability, EI calculated using the intake-balance technique (EE and change in ES) was also highly variable, with moderate correlations between gold-standard and commercial devices, yet not statistically equivalent.

EE estimated from the Fitbit Alta at Time 1 showed good agreement with DLW as evaluated using multiple statistical techniques (2-sided t test $P = 0.79$, $r_s = 0.82$, equivalent); however, the findings from Time 2 were somewhat mixed (2-sided t test $P = 0.40$, $r_s = 0.66$, not equivalent). These findings generally align with previous research, although true comparisons are difficult to make. Research on consumer devices often suffers from "research lag," meaning that validation studies on a given device are not disseminated until

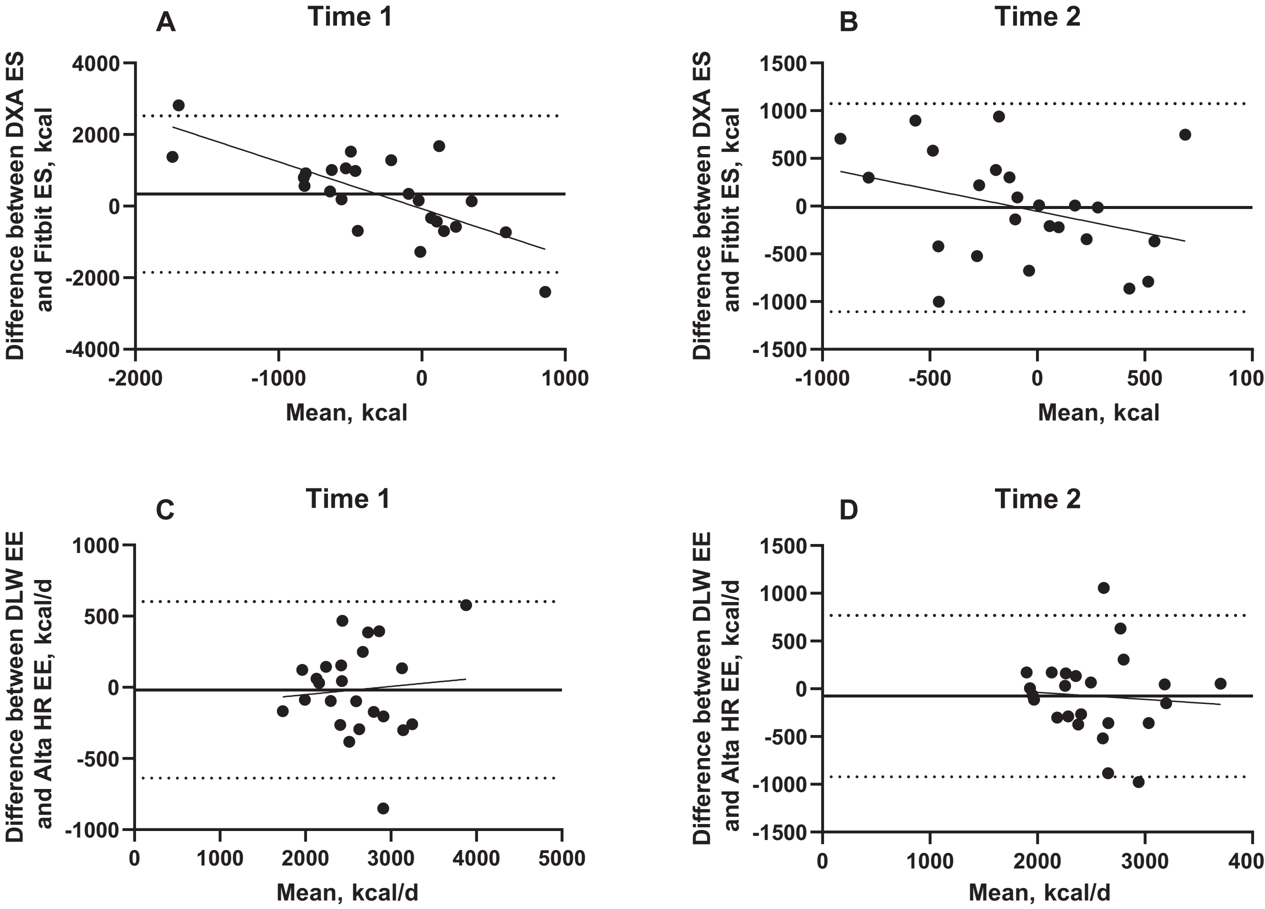

FIGURE 2 Bland–Altman plots for ES (A, B) and EE (C, D) estimates between devices at Time 1 (A, C) and Time 2 (B, D). DLW, doubly labeled water; EE, energy expenditure; ES, energy storage.

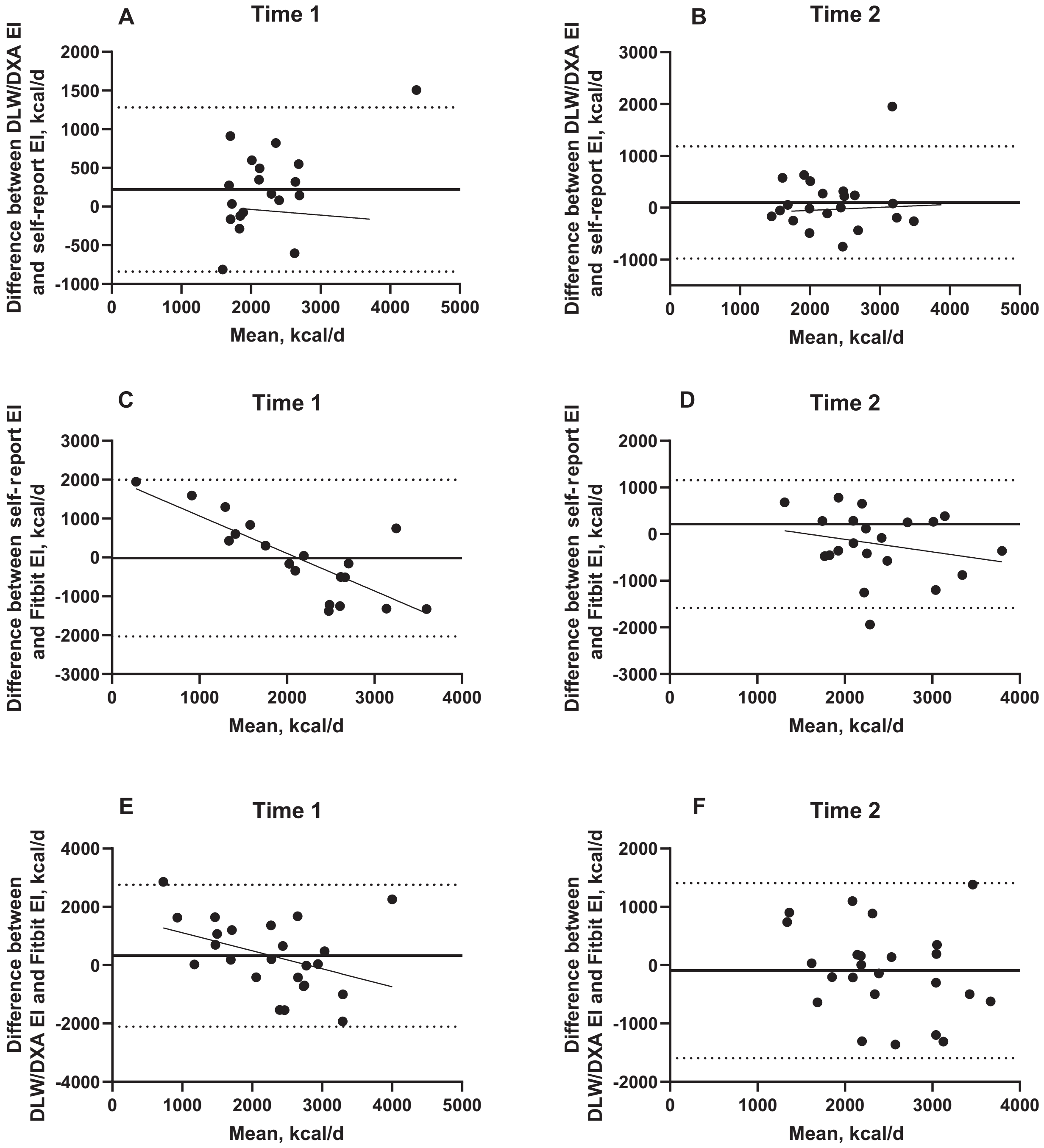

FIGURE 3 Bland–Altman plots for EI estimates between devices and dietary recalls at Time 1 (A, C, E) and Time 2 (B, D, F). DLW, doubly labeled water; EI, energy intake.

after the devices have been replaced by new versions or models (36). Fitbit has historically released multiple new consumer devices or models annually (the Alta was released in 2017 and discontinued in 2019), so we must look historically to provide context for validity and reliability. The Fitbit One, a waist-worn device released in 2012, had an $r = 0.76$ compared with the ActiGraph GTX3 for total daily EE (26). The Fitbit Flex, a wrist-worn monitor released in 2013, had an $r_s = 0.84$ compared with DLW for free-living activities (13). The Fitbit Surge, a wrist-worn device with tri-axial accelerometer released in 2015, had an $r = 0.9$ and an $r^2 = 0.82$ compared with DLW for total daily EE, and was not equivalent at 90% CI of the mean (37). The Fitbit Alta includes a heart rate sensor, although it is not clear if these additional data improve estimates of EE (38).

Assessments of body weight and ES by the Fitbit Aria were generally poor compared with the laboratory scale and DXA, with significant differences observed at time 1 for body weight and FM ($r_s = 0.15$), at time 2 for FM ($r_s = 0.49$), and no equivalence at either time point. Very little rigorous information on the validity of consumer body composition devices in adults exists in the literature, particularly beyond 1 assessment day (39–41). In general, biopolar (feet-to-feet) multifrequency devices have good small bias (r^2 >0.84) but wide limits of agreement (>3 kg for FM) (42–45). We observed

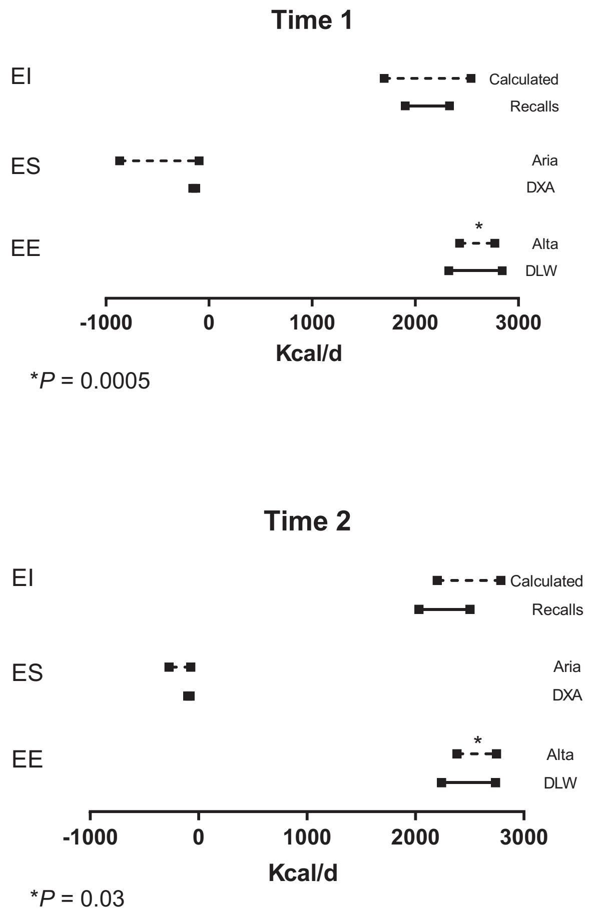

FIGURE 4 Equivalence zones for gold-standard devices and 90% CIs of commercial devices at each time point. Alta, Alta HR activity monitor; Aria, Aria wireless body composition scale; DLW, doubly labeled water; EE, energy expenditure; EI, energy intake; ES, energy storage.

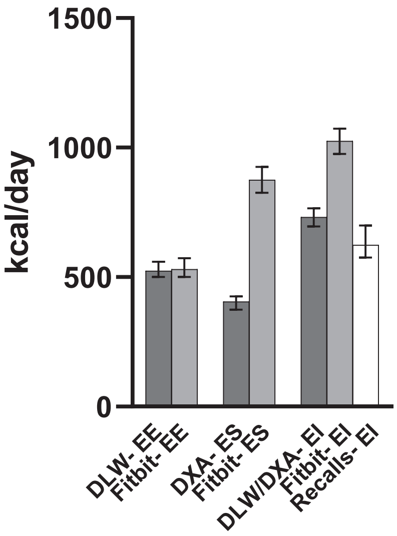

FIGURE 5 Root mean square errors and 95% CIs between device measures and estimates of true measures calculated from linear measurement error modeling. DLW, doubly labeled water; EE, energy expenditure; EI, energy intake; ES, energy storage.

quite poor validity for the Aria during Time 1 which was somewhat improved during Time 2. This may be due to self-calibration after initial setup of the device by the participants, as outlined by the manufacturer: “After movement of the scale…up to two consecutive weigh-ins will then be required before your scale is recalibrated and again displays consistently accurate measurements.” Although other devices such as MRI or computed tomography scans are generally considered “gold standards” to assess body composition, DXA devices generally have been found to have CVs <1.0% for various body compartments when repeated within a given measurement session (46–48).

EI was estimated using the intake-balance technique for both gold-standard and consumer devices and dietician-administered 24-h recall. As with the other energy balance components the results for EI were mixed. There were no differences in estimates observed (P = 0.09–0.94) and some moderate correlations (0.27–0.73), but no statistical equivalence between any of the estimates at any time point. This is not surprising given the variability in assessing EE and ES.

Given the variability that was observed in both the gold-standard and consumer devices in the present study, a useful

TABLE 4 Energy measures compared by 2-sided paired *t* test and by TOST[1]

		TOST		2-sided paired *t* test		
	Time	95% CI	*P* value	95% CI	*P* value	Hedges' *g*
Expenditure						
Fitbit Alta vs. DLW	1	(−93.6, 127.8)	0.0005	(−116.5, 150.7)	0.79	0.03
	2	(−74.6, 226.7)	0.031	(−105.8, 257.9)	0.40	0.15
Storage						
Aria vs. DXA	1	(−727.7, 52.6)	0.93	(−808.5, 133.4)	0.15	− 0.38
	2	(−177.8, 211.3)	0.59	(−218.0, 251.6)	0.88	0.03
Intake						
Fitbit vs. DLW/DXA	1	(−754.8, 114.0)	0.62	(−844.7, 203.9)	0.22	− 0.30
	2	(−175.1, 360.7)	0.18	(−230.6, 416.2)	0.56	0.12
DLW/DXA vs. Recalls	1	(4.59, 435.0)	0.53	(−40.9, 480.5)	0.09	0.27
	2	(−106.0, 309.1)	0.16	(−149.5, 352.6)	0.41	0.15
Fitbit vs. Recalls	1	(−389.5, 428.8)	0.21	(−476.1, 515.4)	0.93	0.02
	2	(−50.5, 475.5)	0.46	(−105.6, 530.6)	0.18	0.29

[1] Comparisons are (Fitbit − gold-standard) or (gold-standard − recall) or (Fitbit − recall). Alta, Alta HR activity monitor; Aria, Aria wireless body composition scale; TOST, two-one-sided-test.

next step would be to use a Bayesian semiparametric approach to evaluate the measurement error of the techniques, with the goal of "calibrating" commercial devices and improving their accuracy (16). This approach has previously been used in other disciplines, including dietary assessment (31, 49), and more recently with self-reported physical activity with some success (50). Given the growth of commercial devices to track activity and body weight, the resurrection of measurement error modeling approaches in the area of energy balance is a logical next step.

There are several strengths of the current study, including the simultaneous use of gold-standard devices and consumer devices under free-living conditions (as opposed to a laboratory setting). In addition, the repeated 14-d protocol allowed for important analysis for repeatability. Limitations include the self-calibration of the consumer body weight and composition scale that occurred during the initial 14-d assessment period which resulted in larger measurement error during that time. In addition, dietary recall data were lost for 5 total participants at Time 1 (2 of them had data at Time 2) owing to laboratory error. However, the completion rate for the remaining 19 participants was quite high (97.5%). Another limitation was that the use of the formula to calculate EI failed to take into account potential uncertainty. Further investigations are needed to incorporate the SEs of the regression coefficients and residual variance when calculating EI.

In conclusion, we observed varying levels of validity and reliability of consumer devices when measuring energy balance during free-living conditions compared to gold-standard devices. Whereas estimates of EE from the wrist-worn consumer devices generally agreed with DLW, estimates for the other energy balance variables (ES and EI) were much more mixed. Given the wide adoption of consumer devices and the potential to inform about population levels of energy balance, future research should identify the measurement error to improve the estimates of EE, ES, and EI.

Acknowledgments

We thank Jennifer Rood from Pennington Biomedical Research Center for her assistance with the DLW assessment. The authors' responsibilities were as follows—RPS, GJW, AMD, and DR: designed the study; RPS: conducted the research, wrote the manuscript, and had primary responsibility for the final content; RPS, H-WY, and DR: analyzed the data; and all authors: read and approved the final manuscript.

Data Availability

Data described in the article, code book, and analytic code will be made available upon request pending application and approval.

References

1. Winkler JT. The fundamental flaw in obesity research. Obes Rev 2005;6(3):199–202.
2. Archer E, Hand GA, Blair SN. Validity of U.S. nutritional surveillance: National Health and Nutrition Examination Survey caloric energy intake data. PLoS One 2013;8(10):e76632.
3. Schoeller DA, Thomas D, Archer E, Heymsfield SB, Blair SN, Goran MI, Hill JO, Atkinson RL, Corkey BE, Foreyt J, et al. Self-report-based estimates of energy intake offer an inadequate basis for scientific conclusions. Am J Clin Nutr 2013;97(6):1413–5.
4. von Helmholtz H. Uber die Erhaltung der Kraft Uber die Erhaltung der Kraft, Ein Physikalische Abhandlung, vorgetragen in der Sitzung der physicalischen Gesellschaft zu Berlin am 23sten Juli 1847. Berlin (Germany): Druck and Verlag von G. Reimer; 1847.
5. Gilmore LA, Ravussin E, Bray GA, Han H, Redman LM. An objective estimate of energy intake during weight gain using the intake-balance method. Am J Clin Nutr 2014;100(3):806–12.
6. de Jonge L, DeLany JP, Nguyen T, Howard J, Hadley EC, Redman LM, Ravussin E. Validation study of energy expenditure and intake during calorie restriction using doubly labeled water and changes in body composition. Am J Clin Nutr 2007;85(1):73–9.
7. Thomas DM, Martin CK, Redman LM, Heymsfield SB, Lettieri S, Levine JA, Bouchard C, Schoeller DA. Effect of dietary adherence on the body weight plateau: a mathematical model incorporating intermittent compliance with energy intake prescription. Am J Clin Nutr 2014;100(3):787–95.
8. Hall KD, Chow CC. Estimating changes in free-living energy intake and its confidence interval. Am J Clin Nutr 2011;94(1):66–74.
9. Sanghvi A, Redman LM, Martin CK, Ravussin E, Hall KD. Validation of an inexpensive and accurate mathematical method to measure long-term changes in free-living energy intake. Am J Clin Nutr 2015;102(2):353–8.
10. Alger K. Wearable technology is revolutionizing fitness. [Internet]. [Accessed 2017 Mar 14]. London (United Kingdom): Raconteur Media; 2014. Available from: https://www.raconteur.net/technology/wearables-are-the-perfect-fit.
11. El-Amrawy F, Nounou MI. Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial? Healthc Inform Res 2015;21(4):315–20.
12. Lee J-M, Kim Y, Welk GJ. Validity of consumer-based physical activity monitors. Med Sci Sports Exerc 2014;46(9):1840–8.
13. Murakami H, Kawakami R, Nakae S, Nakata Y, Ishikawa-Takata K, Tanaka S, Miyachi M. Accuracy of wearable devices for estimating total energy expenditure: comparison with metabolic chamber and doubly labeled water method. JAMA Intern Med 2016;176(5):702–3.
14. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, Hamilton CB, Li LC. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth 2018;6(8):e10527.
15. Shook RP, Hand GA, O'Connor DP, Thomas DM, Hurley TG, Hebert JR, Drenowatz C, Welk GJ, Carriquiry AL, Blair SN. Energy intake derived from an energy balance equation, validated activity monitors, and dual X-ray absorptiometry can provide acceptable caloric intake data among young adults. J Nutr 2018;148(3):490–6.
16. Ries D, Carriquiry A, Shook R. Modeling energy balance while correcting for measurement error via free knot splines. PLoS One 2018;13(8):e0201892.
17. Schoeller DA, Colligan AS, Shriver T, Avak H, Bartok-Olson C. Use of an automated chromium reduction system for hydrogen isotope ratio analysis of physiological fluids applied to doubly labeled water analysis. J Mass Spectrom 2000;35(9):1128–32.
18. Schoeller D, Luke A. Rapid ^{18}O analysis of CO_2 samples by continuous-flow isotope ratio mass spectrometry. J Mass Spectrom 1997;32(12):1332–6.
19. Thompson FE, Suba AF. Dietary assessment methodology. In: Coulston A, Boushey C, Ferruzzi M, editors. Nutrition in the prevention and treatment of disease. 3rd ed. Cambridge (MA): Academic Press; 2013. p. 5–46.
20. Hebert JR, Ebbeling CB, Matthews CE, Hurley TG, Ma Y, Druker S, Clemow L. Systematic errors in middle-aged women's estimates of energy intake: comparing three self-report measures to total energy expenditure from doubly labeled water. Ann Epidemiol 2002;12(8):577–86.
21. Dwyer J, Ellwood K, Leader NP, Moshfegh AJ, Johnson CL. Integration of the Continuing Survey of Food Intakes by Individuals and the National Health and Nutrition Examination Survey. J Am Diet Assoc 2001;101(10):1142–3.
22. Posner BM, Smigelski C, Duggal A, Morgan JL, Cobb J, Cupples LA. Validation of two-dimensional models for estimation of portion size in nutrition research. J Am Diet Assoc 1992;92(6):738–41.
23. Thomas DM, Martin C, Heymsfield S, Redman L, Schoeller D, Levine J. A simple model predicting individual weight change in humans. J Biol Dyn 2011;5(6):579–99.

24. Thomas DM, Bouchard C, Church T, Slentz C, Kraus WE, Redman LM, Martin CK, Silva AM, Vossen M, Westerterp K, et al. Why do individuals not lose more weight from an exercise intervention at a defined dose? An energy balance analysis. Obes Rev 2012;13(10):835–47.

25. Thomas DM, Schoeller DA, Redman LA, Martin CK, Levine JA, Heymsfield SB. A computational model to determine energy intake during weight loss. Am J Clin Nutr 2010;92(6):1326–31.

26. Ferguson T, Rowlands AV, Olds T, Maher C. The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. Int J Behav Nutr Phys Act 2015;12(1):42.

27. Ludbrook J. Confidence in Altman–Bland plots: a critical review of the method of differences. Clin Exp Pharmacol Physiol 2010;37(2):143–9.

28. Calabro MA, Kim Y, Franke WD, Stewart JM, Welk GJ. Objective and subjective measurement of energy expenditure in older adults: a doubly labeled water study. Eur J Clin Nutr 2015;69(7):850–5.

29. Dixon PM, Saint-Maurice PF, Kim Y, Hibbing P, Bai Y, Welk GJ. A primer on the use of equivalence testing for evaluating measurement agreement. Med Sci Sports Exerc 2018;50(4):837–45.

30. Dixon PM, Pechmann JH. A statistical test to show negligible trend. Ecology 2005;86(7):1751–6.

31. Nusser SM, Beyler NK, Welk GJ, Carriquiry AL, Fuller WA, King BM. Modeling errors in physical activity recall data. J Phys Act Health 2012;9(Suppl 1):S56–67.

32. Gamer M, Lemon J, Fellows I, Singh P. Package 'irr'. Various coefficients of interrater reliability and agreement. 2012.

33. Robinson A. Equivalence: provides tests and graphics for assessing tests of equivalence. R package version 07 2016;2.

34. Stan Development Team. RStan: the R interface to Stan. R package version 2016;2(1).

35. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6(4):284–90.

36. Bunn JA, Navalta JW, Fountaine CJ, Reece JD. Current state of commercial wearable technology in physical activity monitoring 2015–2017. Int J Exerc Sci 2018;11(7):503–15.

37. Siddall AG, Powell SD, Needham-Beck SC, Edwards VC, Thompson JES, Kefyalew SS, Singh PA, Orford ER, Venables MC, Jackson S, et al. Validity of energy expenditure estimation methods during 10 days of military training. Scand J Med Sci Sports 2019;29(9):1313–21.

38. Montoye AHK, Vusich J, Mitrzyk J, Wiersma M. Heart rate alters, but does not improve, calorie predictions in Fitbit activity monitors. J Meas Phys Behav 2018;1(1):9.

39. Hood KM, Marr C, Kirk-Sorrow J, Farmer J, Lee CM, Kern M, Bagley JR. Validity and reliability of a Wi-Fi smart scale to estimate body composition. Health Technol 2019;9(5):839–46.

40. Kabiri LS, Hernandez DC, Mitchell K. Reliability, validity, and diagnostic value of a pediatric bioelectrical impedance analysis scale. Child Obes 2015;11(5):650–5.

41. Shaffer JA, Diaz K, Alcántara C, Edmondson D, Krupka DJ, Chaplin WF, Davidson KW. An inexpensive device for monitoring patients' weights via automated hovering. Int J Cardiol 2014;172(2): e263–4.

42. Bosy-Westphal A, Later W, Hitze B, Sato T, Kossel E, Glüer CC, Heller M, Müller MJ. Accuracy of bioelectrical impedance consumer devices for measurement of body composition in comparison to whole body magnetic resonance imaging and dual X-ray absorptiometry. Obesity Facts 2008;1(6):319–24.

43. Heymsfield SB, Kim JY, Bhagat YA, Zheng J, Insoo K, Ahyoung C, Seongwook J, Jaegeol C. Mobile evaluation of human energy balance and weight control: potential for future developments. Annu Int Conf IEEE Eng Med Biol Soc 2015;2015: 8201–4.

44. Jackson AA, Johnson M, Durkin K, Wootton S. Body composition assessment in nutrition research: value of BIA technology. Eur J Clin Nutr 2013;67(S1):S71–8.

45. Ward LC. Human body composition: yesterday, today, and tomorrow. Eur J Clin Nutr 2018;72(9):1201–7.

46. Hind K, Oldroyd B, Truscott JG. *In vivo* precision of the GE Lunar iDXA densitometer for the measurement of total body composition and fat distribution in adults. Eur J Clin Nutr 2011;65(1): 140–2.

47. Leonard CM, Roza MA, Barr RD, Webber CE. Reproducibility of DXA measurements of bone mineral density and body composition in children. Pediatr Radiol 2009;39(2):148–54.

48. Baracos V, Caserotti P, Earthman CP, Fields D, Gallagher D, Hall KD, Heymsfield SB, Müller MJ, Rosen AN, Pichard C, et al. Advances in the science and application of body composition measurement. N J Parenter Enteral Nutr 2012;36(1):96–107.

49. Nusser SM, Carriquiry AL, Dodd KW, Fuller WA. A semiparametric transformation approach to estimating usual daily intake distributions. J Am Statist Assoc 1996;91(436):1440–9.

50. Welk GJ, Kim Y, Stanfill B, Osthus D, Calabro M, Nusser S, Carriquiry A. Validity of 24-h physical activity recall: physical activity measurement survey. Med Sci Sports Exerc 2014;46(10): 2014–24.