Validity of 24-h Physical Activity Recall: Physical Activity Measurement Survey

Gregory J. Welk[1], **Youngwon Kim**[1], **Bryan Stanfill**[2], **David A. Osthus**[2], **Andres M. Calabro**[1], **Sarah M. Nusser**[2], and **Alicia Carriquiry**[2]

[1]Department of Kinesiology, Iowa State University, Ames, IA

[2]Department of Statistics, Iowa State University, Ames, IA

Abstract

Purpose—The primary purpose of this study was to evaluate the validity of an interviewer-administered, 24-hour physical activity recall (PAR) compared to the SenseWear Armband (SWA) for estimation of energy expenditure (EE) and moderate-to-vigorous physical activity (MVPA) in a representative sample of adults. A secondary goal was to compare measurement errors for various demographic sub-groups (gender, age and weight status).

Methods—A sample of 1347 adults (20–71yrs; 786 females) wore an SWA for a single day and then completed a PAR recalling that previous day's physical activity. The participants each performed two trials on two randomly selected days across a 2 year time span. The EE and MVPA values for each participant were averaged across the two days. Group-level and individual-level agreement were evaluated using 95% equivalence testing and mean absolute percent error (MAPE), respectively. Results were further examined for sub-groups by gender, age and body mass index (BMI).

Results—The PAR yielded equivalent estimates of EE (compared to the SWA) for almost all demographic subgroups but none of the comparisons for MVPA were equivalent. Smaller MAPE values were observed for EE (ranges from 10.3% to 15.0%) than for MVPA (ranges from 68.6% to 269.5%), across all comparisons. The PAR yielded underestimates of MVPA for younger, less obese people but overestimates for older, more obese people.

Conclusions—For EE measurement, the PAR demonstrated good agreement relative to the SWA. However, the use of PAR may result in biased estimates of MVPA both at the group and individual level in adults.

Keywords

validity; self-report; objective measurement; measurement error; adult

Corresponding Author: Name: Gregory J. Welk[1], Mailing Address: Department of Kinesiology, Iowa State University, 257 Forker Building, Ames, IA, 50011, Telephone: 1 - 515 - 294 - 3583, Fax: 1 - 515 - 294 - 8740, gwelk@iastate.edu.

All the authors declare no conflicts of interest.

Supplemental Digital Content 1. tiff

Introduction

Accurate assessments of physical activity (PA) are critical for advancing many important public health questions (40). Emphasis in the field has been on improving the utility of objective monitoring devices and technology-based methods (34). Much less attention has been given to more practical and cost-effective solutions such as improving the accuracy and utility of self-report instruments (6, 35). In addition to being easier and less expensive to use in larger studies (18), self-report measures provide valuable information about the context of PA (e.g. type, purpose, location) and this is critical for understanding the underlying social and environmental prompts influencing behavior (35). Self-report measures also provide unique insights about the relative or "subjective" ratings of PA by participants – an important distinction that is reflected in the U.S. Physical Activity Guidelines (37). Surprisingly, few studies have systematically examined variability in the reporting of PA or the nature and sources of error in self-report measures in a systematic way.

Several recent review papers have reported on the psychometric properties of various physical activity questionnaires in adults (15, 19, 39) but there is considerable variability in the results with regard to age, ethnicity, population size, regions, domains of physical activity and unit of measurement. Results also vary greatly depending on the quality of the underlying criterion measure (e.g. accelerometers, indirect calorimetry, doubly labeled water, pedometers, heart rate monitors, or other forms of self-report measures). These factors have made it difficult to draw any strong conclusions about the reliability and validity of self-report instruments.

The most commonly cited limitations of self-report measures are that they are prone to inaccurate or incomplete recall and/or respondent bias.(1, 13, 24) However, the type and degree of error and bias are directly related to the type of form used and the duration of data being recalled. Time-use diaries are more resistant to recall bias since a person reports or codes their behavior as it occurs (36). Studies have documented advantages of time-use diaries for assessing physical activity behavior (17, 38) but the burden on participants can be high. An alternative to a single-day, time-use diary is a retrospective, single-day recall. While still prone to error, the time frame is short enough that participants are able to accurately recall duration of active and sedentary periods in a reasonably objective way (36). Previous work by our team evaluated the utility of an interviewer-administered 24-hour Physical Activity Recall (PAR) against temporally matched data from SenseWear Armband (SWA) monitors (7). We reported moderate to high correlations ($r = 0.57$) for reported levels of moderate to vigorous physical activity (MVPA) and high correlations ($r = 0.88$) for estimates of energy expenditure (EE). The estimates of group-level EE were within 100 kcal of the estimates from the Sensewear (~ 5%–6% error) and these differences were not statistically or practically significant (effect sizes of < 0.20). A unique contribution of the study (7) is that we observed no differences in validity when used with or without a behavioral log designed to remind participants of their behavior. This suggests that participants can provide a reasonably accurate recall of their previous day's activity using this type of interviewer-administered tool. The detail and context available in the PAR also make it an ideal tool to better understand perceptions and reporting of physical activity behaviors in adults.

We recently developed and tested a refined version of the PAR instrument and have used it as the primary outcome measure in a large-scale study designed specifically to study and improve self-report measures of PA. The PAR protocol was used as the primary outcome measure in a 5 year NIH-funded project (Physical Activity Measurement Study (PAMS) - R01 HL91024-01A1) designed to establish and evaluate measurement error models that can be used to improve the utility of self-report measures. The PAR was selected because of its established validity but also because it provides an ideal complement to the established 24h diet recall time instrument currently used in the National Health and Nutrition Examination Survey (NHANES). Previous research has shown that the 24h diet recall has stronger associations with gold standard dietary measures than other longer recall formats (32). Research has also demonstrated the ability to create measurement error models with this short term recall format (8, 12, 29). Our goal was to develop parallel measurement error models to improve the accuracy of physical activity self-reports so the PAR was the logical choice. Preliminary models have already been described (28) but, before these models can be finalized, it was deemed important to directly quantify the type and extent of measurement error in the sample. Therefore, the purposes of this study were 1) to evaluate agreement between PAR and an objective method for estimating EE and MVPA, and 2) to quantify the amount and patterns of measurement errors for various demographic sub-groups in a representative sample of adults.

Method

Study design

The PAMS study used a stratified sampling plan to recruit a representative sample of adults (ages 20–75yrs) from four target counties in the state of Iowa, U.S.A. The counties varied in size (2 urban and 2 rural). Participants were recruited through telephone screening using a sample purchased from Survey Sampling International. A stratified sampling plan was used to increase representation of minorities. Each of the four counties was divided into two co-strata based on higher or lower percentages of Blacks and Hispanics (using tract-level demographic information from the 2000 Census). This allocation across county sub-strata was used to balance recruitment of Blacks and Hispanics across the counties, while keeping the variation in the weights down to the extent possible. Participants were required to be between 20–75 years of age, able to walk, and able to complete both a telephone and written paper survey in either English or Spanish. Recruitment and study procedures were approved by the local Institutional Review Board and written consent was obtained from all participants.

Instruments

SenseWear Armband Mini (SWA)—The SWA (BodyMedia, Inc., Pittsburgh, PA) is a non-invasive, multi-sensor based activity monitor worn on the back of an upper arm. It estimates time spent in different intensities of physical activity (and energy expenditure) based on a combination of multiple sensors (detecting heat flux, galvanic skin response, skin temperature, near body temperature) and a tri-axial accelerometer (capturing body movement). The SWA has been shown to provide valid estimates of EE (21) and MVPA (4) in previous studies. The ease of use and the ability to directly measure non-wear time also

make it particularly attractive for field-based monitoring studies. Based on these factors, the SWA was selected as an appropriate criterion measure to evaluate the PAR. The SWA data were processed using the latest version of software v8.0 (algorithm v5.2).

24 Physical Activity Recall (PAR)—The PAR is a structured, interviewer-administered PA assessment instrument that provides details about the type and intensity of occupational, household and leisure-time PA performed in the past day (7, 25). The PAR interview protocol utilizes a segmented day approach to capture details about the type and intensity of physical activity in 3 time periods (morning, afternoon and evening). The PA data were then converted to estimates of EE using established MET codes from the Compendium of PA (3). The present study used a phone-based methodology that necessitated modifications in the protocol and simplifications to the list of available activity codes. The refinement of the PAR protocol followed an iterative process based on results from a formalized cognitive interview process. The reduced list of activity codes was needed to streamline the selection of codes for the interviewers and to avoid redundancies in the Compendium codes (3). The final database included a list of 270 activity codes with clearly defined descriptions and associated METs values. Details on the cognitive interviews and the final PAR protocol (and PA codes) are available in technical reports and are available by request from authors.

Data Collection

The data for the PAMS project were collected by an experienced and trained research group managed through the Social and Behavioral Research Services group at our Institution. The team and associated field staff collected data over a continuous 2 year period (24 months; 8 3-month quarters) to minimize variability due to seasonality, weather patterns, or other unforeseen circumstances. The protocol included a screening interview to determine eligibility and to recruit a randomly selected adult to participate. Participants were asked to complete two separate 1-day sessions of monitoring (referred to as Trial 1 and Trial 2 hereafter) – each of which included wearing the SWA for a 24-hour period and completing a PAR the following day to recall the activities performed. Members of the field staff came to each participant's residence to initialize the monitor and to provide instructions about the study. Participants were asked to wear the monitor for the full 24 hour period (except for showers) but they were provided with a log to record duration and type of activity performed if the monitor was removed. The staff member picked up and downloaded the monitor the next day and scheduled the follow-up PAR interview.

The PAR interviews were conducted by a team of trained interviewers using a computer – assisted telephone interview (CATI) system programmed using the *Blaise* Trigram methodology. The participants were asked to recall his/her physical activity level on the day they wore the monitor. The day was divided into four, 6-hour blocks starting from the midnight, and participants were asked to report only activities performed for 5 minutes or greater. The interviewers would select the named activity from a computerized list of activities (based on a reduced set of activities from the Compendium of Physical Activity (3)), and then record the location, purpose, and number of minutes for each activity. The interviewers used a series of semi-structured probes to prompt recall and to facilitate accurate recall of the day by the participant. All PAR interviews were conducted under the

supervision of project staff and Spanish-speaking interviewers were available to handle calls that could not be completed in English. The interviewers tracked total reported minutes to ensure that participants accounted for the full 1440 minutes in the day. The PAR interviews took an average of 20min to complete (range from 12 to 45min).

Data Processing

The data from the SWA and PAR were processed to facilitate direct comparisons of both EE and time spent in PA. An advantage of the SWA is that it automatically detects non-wear time and this feature greatly facilitates data processing and screening. The software fills non-wear time periods with estimates of resting EE but this would likely lead to some underestimation of EE since these periods are more likely to be at least of light intensity. Therefore, cases with observed non-wear time were corrected using data from the participant logs. Specifically, we used established MET values from the Compendium of Physical Activities to adjust the EE estimates (and corresponding PA estimates) for all logged activities. Periods of non-wear time that were not logged were filled using a MET value of 1.5 to correspond with light PA. The SWA also directly tracks minutes spent in moderate PA (3–6 METS) and vigorous PA (> 6 METS) and these allocations were combined to obtain an indicator of total minutes of MVPA per day. Minutes of moderate or vigorous PA reported on the log (when the monitor was not worn) were also added to the allocations for MVPA.

The data from the PAR were processed to obtain parallel estimates of total daily EE and total minutes of MVPA. Each PAR Code in the dataset was assigned a corresponding MET code based on the Compendium of Physical Activities and converted to estimates of EE using standard MET-based formulas: $EE = ((METs \times weight (kg))/60) \times minutes$. The EE estimates and minutes were aggregated by participant and Trial and the accumulated minutes were checked to confirm that all participants had 1440 minutes of coded data. The EE values were then rescaled using World Health Organization (WHO) estimates of resting EE (REE) (14). This was an important modification in the coding of the PAR since the SWA also uses the WHO values as the basis for determinations of REE in estimates from the SWA.

Statistical Analyses

Weighted means, mean errors (ME) and correlations were calculated to account for the complex sampling design of this study. Survey weights were used to account for the complex sampling design of this study, as individuals from co-strata with higher percentages of Blacks and Hispanics were over sampled. Thus, survey weighted means, mean errors (ME) and correlations were calculated, as they are more representative of the underlying population than calculations based on the unweighted sample. Standard errors were calculated using a customized jackknife procedure, following that of Section 4.2.2 of Fuller (16).

Group-level agreement between the PAR and SWA was evaluated using 95% equivalence testing (11). In traditional hypothesis testing of a difference, the null hypothesis is that two methods are the same; however, failure to reject the null does not necessarily imply that

there is equivalence between the two measures (only that there is no evidence that they are the same). However, with the 95% equivalence testing approach, the null hypothesis is that there is a difference between the PAR and SWA. This, in turn, enables the corresponding alternative hypothesis to specify that there is no difference (i.e. equivalence). To test this, we used two one-sided tests to determine whether the difference between the PAR and SWA fit within a specified zone of equivalence (which was defined as ± 10% of the means of the SWA herein). Rejecting these two one-sided tests corresponds to 95% equivalence (i.e. loss of 5% confidence at each of the two ends) (11). To satisfy the assumption of normality and constant variance with our sample it was necessary to use a log scale. Therefore, we specifically tested if the 90% CIs of group-mean differences between the PAR and SWA (on a log scale) were completely included between logs of 0.9 and 1.1 (using an alpha level of 5% for the tests). In this case, a log of 1 indicates equivalence (i.e. a difference of 0), and the range in log from 0.9 to 1.1 corresponds to specify the equivalence zone (i.e. ± 10% of the means of the SWA.

To investigate individual-level agreement, mean absolute percent errors (MAPE) and Bland-Altman plots were used. Traditional Bland-Altman plots (average on x-axis and difference on y-axis) showed evidence of relationships and non-uniform differences in the average and difference between the PAR and SWA. Therefore, instead of plotting differences against averages, ratios (i.e. PAR over SWA) in relation to averages were presented, and a regression approach was carried out to illustrate the non-uniform differences. With this approach, a regression line with an intercept of 1 and slope of 0 would indicate the two instruments agree with one another (on average) across averages. A significant deviation from that trend, as evidenced by 95% limit of agreement, for example, would indicate that the two methods do not agree. To estimate the slope and intercept parameters for the regression approach, weighted least-squares were used. Standard error estimates were also computed using the same jackknife procedure mentioned above.

In order to determine interaction effects among gender, age, and BMI for differences in EE and MVPA estimates, a repeated measure ANOVA was fit to the differences on the log-scale. Again, the log-scale measurements were used to satisfy the assumptions for ANOVA. Type III sums of squares were used to determine which variables and interactions needed to be included in the model. The resulting interaction effects were presented and illustrated on heat maps. All of the analyses were performed using R for Windows.

Results

A total of 1648 participants completed the screening process and were eligible to participate, but a sample of 1501 were officially enrolled in the study. The majority of the missing cases (n = 141, 96%) were due to participants changing their mind about participating, but others were lost due to death (n=1), relocation (n = 3) and / or becoming pregnant (n = 2). A total of 2981 data records were obtained in the study (Trial 1: n = 1501; Trial 2: n = 1480), but 149 cases were not suitable for analyses. Some cases were lost due to missing PAR data (n = 13), outliers on the PAR (n = 7) or to staff data entry errors with the PAR (n = 18). Additional cases were lost due to blank SWA files (n = 32), non-compliance in monitoring (n = 2), data transfer problems (n = 10) and staff coding errors (n = 3). A final set of SWA

files (n = 64) could not be re-processed using the latest version of the algorithms and were therefore excluded. The reduced set of 2832 records had 1468 unique ID's but only 1347 had complete data on both trials. This was an important measurement consideration in the study so the final data set was restricted to the 2694 replicate records from this sample of 1347 participants.

The overall compliance with the 24 hour monitoring protocol was very high with average wear time of 98.6%. Approximately, 25% of participants wore the monitor for all 24 hours, another 60% removed the monitor for less than 30 minutes (presumably for shower or changing clothes) and a relatively small number of cases (15%) had the monitor off for 30 minutes or more. Sensitivity analyses was conducted to determine if the distributions or patterns varied when the missing data were filled using the logs or the default estimates of light activity (1.5 METS) and no issues were noted. Therefore, it was determined to retain the estimates from all participants.

For the present analyses, the final sample was restricted to the 1347 participants that had complete data on both instruments on both trials. The sample included approximately 7% Black and 5% Hispanic participants (values which are similar or higher than the overall distribution in the state). There was an even gender distribution (49% female) and there was reasonable distribution across the four age groups (20–29, 30–39, 40–49, and 50–71yrs) with the largest fraction of participants (~ 35%) in the 40–49 year range (see Table 1). Approximately 75% of the sample was overweight or obese but the percentages were somewhat lower for the 20–29 age group (55%) and 30–39 age group (68%) than for the 40–49 age group (77%) and 50–71 age group (77%).

Table 2 shows the descriptive statistics for both the PAR and SWA as well as group level calculations of error (both mean error) and MAPE as well as the correlation (rho) between estimates. The top panels show estimates of EE (kcal/day) while the bottom panels report estimates of time spent in MVPA (minutes/day). The results reveal the expected demographic patterns with males having higher reported and observed MVPA and higher reported and observed estimates of EE. The estimates of time spent in MVPA was high (males: 168min; females 95min) – at least compared to data on MVPA levels of the US adults reported in other studies (33). The time spent in MVPA followed expected age related patterns (i.e. lower values with advanced age) for both the PAR and the SWA and this was manifested in similar age-related declines in estimates of EE. The objective SWA data revealed subgroup differences in observed time spent in MVPA, with normal weight and overweight participants having higher levels of MVPA than obese participants. However, similar estimates of MVPA were observed with the subjective PAR data.

The error and MAPE values in Table 2 show the overall accuracy of the group level estimates. It is important to note that 10 of the 1347 participants reported 0 min of MVPA (which would cause undefined MAPE values). To resolve this issue we calculated MAPE values on the sub-sample of 1337 with reported MVPA (i.e. 1347-10). We used the same sub-sample to calculate MAPE for EE to facilitate direct comparison with MAPE for MVPA. Overall, the PAR appeared to underestimate EE values (ranges from −315.3 to −129.9kcals/day relative to the SWA. For EE estimates, the error was larger in males (ME =

−329.7 kcal) than in females (ME = −129.9 kcal) but the differences in MAPE values were not as large. Values were similar across age groups (~ 11–12%) with the exception of the young sample which had a MAPE of approximately 15%. There was no large difference in MAPE values across the weight categories. The overall accuracy of the group level estimates was examined more quantitatively using equivalence testing and these results are presented in Figure 1. As described, with 95% equivalence testing, the 90% confidence intervals of the PAR are compared to a predefined interval of values which in this case was set at ±10% of the mean of the SWA. Thus, 95% confidence intervals that are completely falling within this criterion (represented as two solid vertical lines at −0.1 and 0.1 on the x-axis in Figure 1) were considered to be equivalent. As shown in the figure, the agreement was fairly good for estimation of EE as the majority of the demographic comparisons (i.e. for all, female, ages 30–39yrs, 40–49yrs, 50–71yrs, overweight and obese group) met the criteria for equivalence. However, the estimates of MVPA were not equivalent to the SWA for any of the comparisons.

The group level estimates in Table 2 and Figure 1 provide an indication of overall agreement but they provide little information about individual level agreement or bias. Bland Altman plots were created to illustrate the overall distribution and patterns of error across the range of activity (and energy expenditure) levels. They cannot, however, be used to identify in what respects (mean or variance) these measures agree or disagree (5). That is, the pattern exhibited by the data in Figure 2 may be due to the fact that the mean of the two measures differ, the variances of the two measures differ, or both. Overall, the distribution of error for the EE estimates was evenly distributed and there was little indication of bias; however, the plot for MVPA reveals a wider range of 95% limits of agreement in comparison with the plot for EE, and larger error for less active individuals in comparison with for more active individuals (Figure 2). A key goal in the study was to examine the nature of error across different demographic variables. Therefore, Bland Altman plots were presented for each of the demographic variables for EE and MVPA estimation, and showed patterns of systematic bias (analogous to the overall comparison) for MVPA that larger ranges of 95% limits of agreement and larger errors for less active individuals were identified (see Figure, Supplemental Digital Content 1, which illustrates patterns of systematic bias for EE and MVPA estimates by demographic variables on Bland Altman plots). No particular patterns of bias were observed for Bland Altman plots for EE estimation.

Table 3 provides a summary of the weighted slopes and intercepts for the various Bland Altman plots relating the PAR to the SWA. These estimates are included to compare the average behavior of these instruments across the range of observed activities but cannot be interpreted directly as bias due to the presence of measurement error (5). The table provides values for both EE and MVPA outcomes using the combined data as well as when segmented by the key demographic variables (gender, age group and BMI group). Across all the comparison groups, EE showed better regression coefficients (intercept closer to 1 and slope closer to 0) than MVPA.

The Bland Altman plots hint at bias but it is not possible with this type of image to determine whether the bias varied across different segments of the population. To address this limitation, we created "heat maps" which reflect interactions among gender, age, and

BMI on estimations of both EE and MVPA (See Figure 3). The white zones reflect areas where the SWA and PAR are the same (on average). The blue indicate values from the PAR that are smaller than SWA while the red zones indicate values from the PAR that are larger than the SWA. For the EE plot (top), it is evident that as BMI increases, there is a greater tendency for overestimation on the PAR relative to the SWA. This pattern was evident for both females and males. The plot for MVPA (bottom) is more complicated. For females, younger individuals with lower BMI (lower-left corner) and older individuals with high BMI (upper-right corner) tend to under-report their PA level relative to the SWA, while the rest of the females over-report their PA levels. For males, however, the opposite is true. More specifically, older males with low BMI (upper-left corner) and young males with high BMI (lower-right corner) tend to under-report their PA levels while the rest of the males over-reporting their PA.

Discussion

A better understanding of measurement error is essential to advance research on PA assessment. Self-report instruments (like any measurement technique) are subject to random measurement error (as well as bias), but these sources of error can potentially be prevented and/or corrected once they are better understood. An NIH consensus conference was recently convened to help advance research on self-report measures of PA (6) and specific recommendations were proposed to better understand (and quantify) sources of error in self-report methods (2, 28). The present study followed recommended validation practices (23) to specifically advance knowledge about measurement error in the PAR instrument, a commonly used recall format. An advantage of the PAR is that the time frame is proximal enough for participants to be able to accurately recall their activity behaviors (as well as the associated context and setting). Another advantage is that the PAR format parallels the commonly used 24 hour diet recall format widely used in research and public health surveillance. Researchers in nutrition have developed robust measurement error models to improve the utility of food recalls (8, 12, 29) and parallel models would offer the same potential to improve the accuracy and utility of physical activity self-report (28).

The overall goal of the PAMS project was to develop measurement error models for the PAR that would adjust for (and correct) underlying biases (i.e. recall bias (13), social desirability bias(1)) in PA self-report data. Before this can be done it was important to better understand the sources and distributions of error in different segments of the population. The direct temporal coupling of the data in the study made it possible to make direct comparison between the observed data from the SWA and the reported data from the PAR. Separate evaluations were conducted for both PA and EE since they reflect different constructs and are influenced by different assumptions. In particular, the estimates of PA from the PAR are based on subjective reports about the relative intensity of various bouts of PA while the estimate from the SWA reflects the observed (objective) pattern of monitored data from the multi-sensor SWA. The conversion to EE for both instruments is also based on different assumptions. With the SWA, the pattern of data is converted to EE using complex pattern recognition algorithms which produce estimates of EE for each of the various categories of PA (i.e. sedentary, light, moderate, vigorous) that are detected. With the PAR, the estimates are predicated on assumptions associated with the use of METs as well as the accuracy of

estimates from the established Compendium of Physical Activities (3). The sources of error contributed by these conversions are likely considerable, but they are dwarfed by the various sources of individual variability as well as by random error.

Considering these issues and assumptions, the overall agreement between the PAR and the SWA in the study was quite impressive. The participants, on average, slightly over-estimated their involvement in MVPA relative to the SWA but the estimates were within 10–20 minutes per day (< 10% of daily totals). The estimates of EE were also within about 10% of daily totals with error ranging from 100–300 kcal/day (depending on the segment of the sample). However, an interesting observation from the analyses is that the PAR generally resulted in overestimation of MVPA, but an underestimation of EE relative to the SWA. This pattern is somewhat counterintuitive as one would expect over-reporting of PA to lead to over-estimates of EE. A number of factors may influence these relationships. It is possible that the MET estimates used to convert the PAR codes into estimates of EE contribute to the underestimation of EE. It is also possible that the SWA may underestimate EE systematically or even with bias. The SWA was chosen for the PAMS study because it has been shown to be more accurate than other competing, accelerometry-based monitors; however, it cannot be considered a true gold standard since it is also "estimating" PA and EE. This conundrum (the lack of a true field-based criterion measure) is what makes it particularly difficult to fully understand the error in self-report measures.

Some insights can be gained with comparisons from a previous study conducted by our team (7) using a slightly different version of the PAR protocol (25). In that study (7), we observed non significant differences between PAR and an earlier version of the SenseWear Armband (SenseWear Armband Pro2) BodyMedia, Inc., Pittsburgh, PA) for estimates of EE, but significant underestimations of MVPA were evident. It is not clear why the participants may have underestimated MVPA but an underestimation would tend to reduce the associated estimate of EE and this may have contributed to the very small differences in estimates of EE observed in the study (7). There was some evidence of systematic bias in this preliminary study (7) with the PAR tending to underestimate (relative to the armband) for participants that were less active but to overestimate (relative to the armband). Despite the differences, the results match those of the present study with higher agreement for EE than for MVPA and considerable individual variability.

A study on an adult version of the Multimedia Activity Recall for Children and Adolescents (MARCA) instrument (17) used a similar computerized, time-based PAR instrument and it was found to have excellent test-reliability and moderate to strong validity coefficients for estimates of EE (when compared to data from an Actigraph accelerometer). However, the small sample size (n = 38) and limited detail about the nature and magnitude of the error make it hard to draw comparisons with the present study. However, results from several other studies provide valuable insights to facilitate interpretation of the findings. A study by Matthews et al. (26) examined the validity of the PAR instrument (relative to the activPAL) in a sample of 88 adults and 81 adolescents. In that study (26), strong positive relationships were observed between the PAR and activPAL for estimates of sedentary time (i.e. regression slopes ranging from 0.80 to 1.13; correlations ranging from 0.68 to 0.77) and time spent in physical activities (i.e. regression slopes ranging from 0.64 to 1.09;

correlations ranging from 0.52 to 0.80). The authors noted that reporting errors inherent in the PAR were smaller than the magnitude of random errors. A study by Slootmaker et al. (31) examined agreement between self-report data and the PAM monitor (PAM, model AM101, PAM B.V., Doorwerth, The Netherlands) by age, sex, education and weight status. The sample included 301 young adults (22–40 years old) and the authors reported "moderate agreement" between self-reported time and objectively measured time spent on MPA. The reported time spent in PA were larger than the observed estimates (MPA: mean difference = 107 +/− 334 min/wk; VPA: mean difference = 169 +/− 250 min/wk) but this varied within the sample. Overweight adults reported significantly more VPA (57 min/wk, $p = 0.04$) than normal weight adults, but this difference was not evident in the accelerometer data. These results are consistent with the observations in our study of systematic over-reporting of MVPA and larger differences for obese individuals. Other studies have reported larger error in estimates of EE (22) and MVPA (i.e. overestimation) (30) and / or poorer agreement (10) in overweight and obese individuals.

While it is logical to label these differences as "error", it is possible (and perhaps likely) that the differences simply reflect differences in perception of PA in overweight or obese individuals. This is an important, and under-studied, distinction considering the importance placed on "relative" intensities of PA in the U.S. Physical Activity Guidelines (37). An objective monitoring device would objectively categorize activities the same for all individuals but a bout of "moderate" intensity activity may (for example) be perceived as "vigorous" for an unfit person or "light" for a highly fit person (20). Few studies have systematically examined possible differences in perception of PA in typical adults but it makes conceptual sense that an obese individual would perceive the same bout of PA as being more intense or tend to report more of it. It is not possible to determine the actual relative intensity of PA without an indication of maximal aerobic capacity, but the main rationale for stratification of our results by gender, age and BMI was to try to examine these differences in a systematic way. In addition to this BMI effect, we anticipated differences due to age since older adults may also perceive a bout of PA differently than young individuals. This difference was not evident in our sample but it is not possible to completely discount an effect of age with the present data. The heat maps shown in Figure 3 display the complexity of error distributions across several variables but it is likely that still other variables such as education, occupation and past PA involvement would contribute to differences in reporting of PA. The overall goal of the PAMS project is to ultimately develop measurement error models to obtain more accurate group-level estimates of PA but the results presented here demonstrate the challenges of obtaining accurate group-level and individual-level estimates of PA (in particular, MVPA) from self-report instruments. A recent study (9) demonstrated good responsiveness for evaluating time spent being sedentary using a specific past day recall of sedentary behavior but it is not possible to directly evaluate responsiveness in the present study or to provide insights about assessments of sedentary behavior. Additional work is needed to evaluate the utility of the PAR for evaluating sedentary time.

Overall, our results support the continued use of the PAR protocol for additional research on PA profiles as well as for broader surveillance applications. Previous studies have found it to be more robust to recall bias (27) and to have better validity than traditional self-report tools

(38). The present results builds on previous work (7) and shows that group level estimates of MVPA and EE were within approximately 10% of monitored values (based on mean percent error). The results also demonstrate reasonable individual accuracy for estimation of EE with MAPE values of approximately 10–12%; however, estimates of MVPA have considerable error (> 68.6% of MAPE). The findings will inform refinement of our measurement error models (28) and this will make it possible to obtain more accurate estimations of PA in population-based research.

The key strengths of the study include the use of a large and representative sample of adults as well as the use of randomly assigned monitoring days across a 2 year period. This feature controls for seasonal variability and allows the data to reflect patterns for typical adults. Key methodological strengths of the study include the use of equivalence tests which are designed to determine if the EE and PA values from the PAR can be considered equivalent to the SWA values. Traditional hypothesis tests in validation studies use difference tests but these can erroneously suggest no evidence of a difference if the data are noisy or if the sample size is small. With an equivalence test, the null hypothesis is flipped to specify that there is a difference and this makes it possible to directly test the equivalence (i.e. rejection of the null hypothesis). A limitation is that there is currently no consensus on what criteria to use to establish an equivalence zone for this type of study. However, we contend that this approach offers conceptual and analytical advantages over traditional difference tests. Other strengths of the study include the tabular reporting of the slopes and intercepts from the Bland Altman plots and the use of heat maps to display the distribution of error across the targeted demographic variables examined in the study. Our results are specific for the PAR but we would expect the findings to generalize to other self-report tools. Another limitation of the study is that we obtained data on only 2 independent 24 hour periods. This prevents the data from being representative of individual behavior but, on a population level, the data can be considered representative of adult PA behaviors due to the random assignment and testing protocols that were used. Overall, the results of the study provide some unique insights about the type, nature and extent of error in self-report measures from the PAR in a representative sample of adults.

Conclusion

The PAR demonstrated good agreement relative to the SWA in estimating EE both at the group and individual level, and did not show any specific pattern of systematic bias. However, large group- and individual-level errors were observed for MVPA estimates. In addition, the distributions of errors varied depending upon the demographic variables in estimating MVPA. Researchers should be aware of the potential large bias inherent in estimating MVPA with the PAR in adults.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Adams SA, Matthews CE, Ebbeling CB, et al. The effect of social desirability and social approval on self-reports of physical activity. Am J Epidemiol. 2005; 161(4):389–398. [PubMed: 15692083]

2. Ainsworth BE, Caspersen CJ, Matthews CE, Masse LC, Baranowski T, Zhu W. Recommendations to improve the accuracy of estimates of physical activity derived from self report. J Phys Act Health. 2012; 9(Suppl 1):S76–S84. [PubMed: 22287451]

3. Ainsworth BE, Haskell WL, Herrmann SD, et al. 2011 Compendium of Physical Activities: a second update of codes and MET values. Med Sci Sports Exerc. 2011; 43(8):1575–1581. [PubMed: 21681120]

4. Berntsen S, Hageberg R, Aandstad A, et al. Validity of physical activity monitors in adults participating in free-living activities. Br J Sports Med. 2010; 44(9):657–664. [PubMed: 18628358]

5. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet. 1995; 346(8982):1085–1087. [PubMed: 7564793]

6. Bowles HR. Measurement of active and sedentary behaviors: closing the gaps in self-report methods. J Phys Act Health. 2012; 9(Suppl 1):S1–S4. [PubMed: 22287442]

7. Calabro MA, Welk GJ, Carriquiry AL, Nusser SM, Beyler NK, Mathews CE. Validation of a computerized 24-hour physical activity recall (24PAR) instrument with pattern-recognition activity monitors. J Phys Act Health. 2009; 6(2):211–220. [PubMed: 19420399]

8. Carriquiry AL. Estimation of usual intake distributions of nutrients and foods. J Nutr. 2003; 133(2): 601s–608s. [PubMed: 12566510]

9. Clark BK, Winkler E, Healy GN, et al. Adults' past-day recall of sedentary time: reliability, validity, and responsiveness. Med Sci Sports Exerc. 2013; 45(6):1198–1207. [PubMed: 23274615]

10. Cust AE, Smith BJ, Chau J, et al. Validity and repeatability of the EPIC physical activity questionnaire: a validation study using accelerometers as an objective measure. Int J Behav Nutr Phys Act. 2008; 5:33. [PubMed: 18513450]

11. Dixon PM, Pechmann JHK. A statistical test to show negligible trend. Ecology. 2005; 86(7):1751–1756.

12. Dodd KW, Guenther PM, Freedman LS, et al. Statistical methods for estimating usual intake of nutrients and foods: A review of the theory. J Am Diet Assoc. 2006; 106(10):1640–1650. [PubMed: 17000197]

13. Durante R, Ainsworth BE. The recall of physical activity: Using a cognitive model of the question answering process. Med Sci Sports Exerc. 1996; 28(10):1282–1291. [PubMed: 8897386]

14. FAO/WHO/UNU. Energy and protein requirements; Who Technical Report Series. 1985. p. 1-206.

15. Forsen L, Loland NW, Vuillemin A, et al. Self-administered physical activity questionnaires for the elderly: a systematic review of measurement properties. Sports Med. 2010; 40(7):601–623. [PubMed: 20545382]

16. Fuller, WA. Sampling Statistics. New York: Wiley; 2009.

17. Gomersall SR, Olds TS, Ridley K. Development and evaluation of an adult use-of-time instrument with an energy expenditure focus. J Sci Med Sport. 2011; 14(2):143–148. [PubMed: 20932797]

18. Haskell WL. Physical activity by self-report: a brief history and future issues. J Phys Act Health. 2012; 9(Suppl 1):S5–S10. [PubMed: 22287448]

19. Helmerhorst HJ, Brage S, Warren J, Besson H, Ekelund U. A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. Int J Behav Nutr Phys Act. 2012; 9:103. [PubMed: 22938557]

20. Howley ET. Type of activity: resistance, aerobic and leisure versus occupational physical activity. Med Sci Sports Exerc. 2001; 33(6):S364–S369. [PubMed: 11427761]

21. Johannsen DL, Calabro MA, Stewart J, Franke W, Rood JC, Welk GJ. Accuracy of armband monitors for measuring daily energy expenditure in healthy adults. Med Sci Sports Exerc. 2010; 42(11):2134–2140. [PubMed: 20386334]

22. Mahabir S, Baer DJ, Giffen C, et al. Comparison of energy expenditure estimates from 4 physical activity questionnaires with doubly labeled water estimates in postmenopausal women. Am J Clin Nutr. 2006; 84(1):230–236. [PubMed: 16825700]

23. Masse LC, de Niet JE. Sources of validity evidence needed with self-report measures of physical activity. J Phys Act Health. 2012; 9(Suppl 1):S44–S55. [PubMed: 22287447]

24. Matthews, C. Use of self-report instruments to assess physical activity. In: Welk, G., editor. Physical Activity Assessments for Health-Related Research. Champaign, IL: Human Kinetics; 2002. p. 107-123.

25. Matthews CE, Ainsworth BE, Hanby C, et al. Development and testing of a short physical activity recall questionnaire. Med Sci Sports Exerc. 2005; 37(6):986–994. [PubMed: 15947724]

26. Matthews CE, Keadle SK, Sampson J, et al. Validation of a previous-day recall measure of active and sedentary behaviors. Med Sci Sports Exerc. 2013; 45(8):1629–1638. [PubMed: 23863547]

27. Matthews CE, Moore SC, George SM, Sampson J, Bowles HR. Improving self-reports of active and sedentary behaviors in large epidemiologic studies. Exerc Sport Sci Rev. 2012; 40(3):118–126. [PubMed: 22653275]

28. 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–S67. [PubMed: 22287449]

29. Nusser SM, Carriquiry AL, Dodd KW, Fuller WA. A semiparametric transformation approach to estimating usual daily intake distributions. J Am Stat Assoc. 1996; 91(436):1440–1449.

30. Oostdam N, van Mechelen W, van Poppel M. Validation and responsiveness of the AQuAA for measuring physical activity in overweight and obese pregnant women. J Sci Med Sport. 2012; 16(5):412–416. [PubMed: 23063355]

31. Slootmaker SM, Schuit AJ, Chinapaw MJ, Seidell JC, van Mechelen W. Disagreement in physical activity assessed by accelerometer and self-report in subgroups of age, gender, education and weight status. Int J Behav Nutr Phys Act. 2009; 6:17. [PubMed: 19320985]

32. Subar AF, Kipnis V, Troiano RP, et al. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN Study. American Journal of Epidemiology. 2003; 158(1):1–13. [PubMed: 12835280]

33. Trinh OT, Nguyen ND, Dibley MJ, Phongsavan P, Bauman AE. The prevalence and correlates of physical inactivity among adults in Ho Chi Minh City. BMC Public Health. 2008; 8:204. [PubMed: 18541020]

34. Troiano RP. A Timely Meeting: Objective Measurement of Physical Activity. Medicine & Science in Sports & Exercise. 2005; 37(Supplement):S487–S489. [PubMed: 16294111]

35. Troiano RP, Pettee Gabriel KK, Welk GJ, Owen N, Sternfeld B. Reported physical activity and sedentary behavior: why do you ask? J Phys Act Health. 2012; 9(Suppl 1):S68–S75. [PubMed: 22287450]

36. Tudor-Locke C, van der Ploeg HP, Bowles HR, et al. Walking behaviours from the 1965–2003 American Heritage Time Use Study (AHTUS). Int J Behav Nutr Phys Act. 2007; 4:45. [PubMed: 17897472]

37. United States. Dept. of Health and Human Services. Vol. ix. Washington, DC: U.S. Dept. of Health and Human Services; 2008. 2008 physical activity guidelines for Americans : be active, healthy, and happy!; p. 61ODPHP publication.

38. van der Ploeg HP, Merom D, Chau JY, Bittman M, Trost SG, Bauman AE. Advances in population surveillance for physical activity and sedentary behavior: reliability and validity of time use surveys. Am J Epidemiol. 2010; 172(10):1199–1206. [PubMed: 20855469]

39. van Poppel MN, Chinapaw MJ, Mokkink LB, van Mechelen W, Terwee CB. Physical activity questionnaires for adults: a systematic review of measurement properties. Sports Med. 2010; 40(7):565–600. [PubMed: 20545381]

40. Welk GJ. Principles of design and analyses for the calibration of accelerometry-based activity monitors. Med Sci Sports Exerc. 2005; 37(11 Suppl):S501–S511. [PubMed: 16294113]
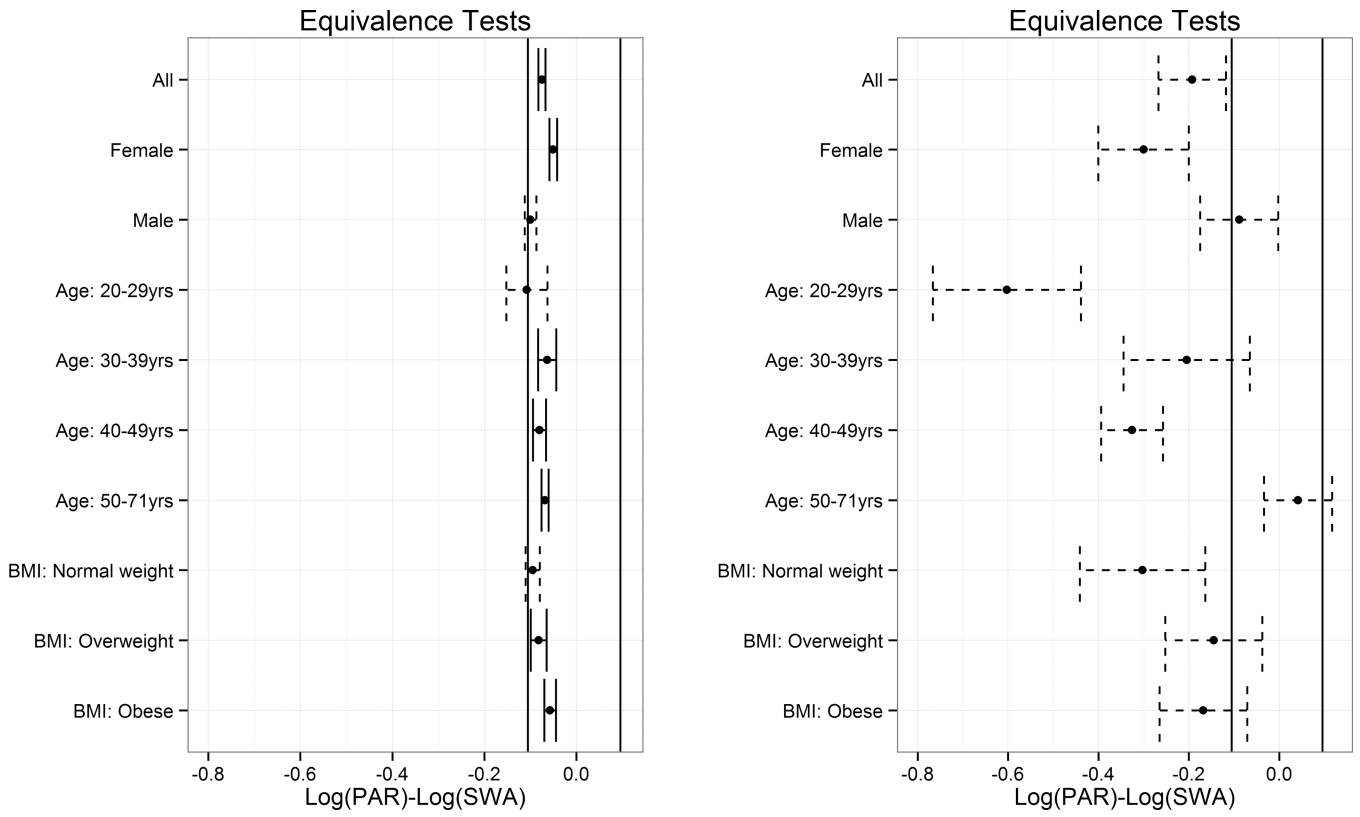
Figure 1.
Results from 95% equivalence testing for agreement between 24-h physical activity recall (PAR) and SenseWear Armband (SWA) for energy expenditure (left) and moderate-to-vigorous physical activity (right).
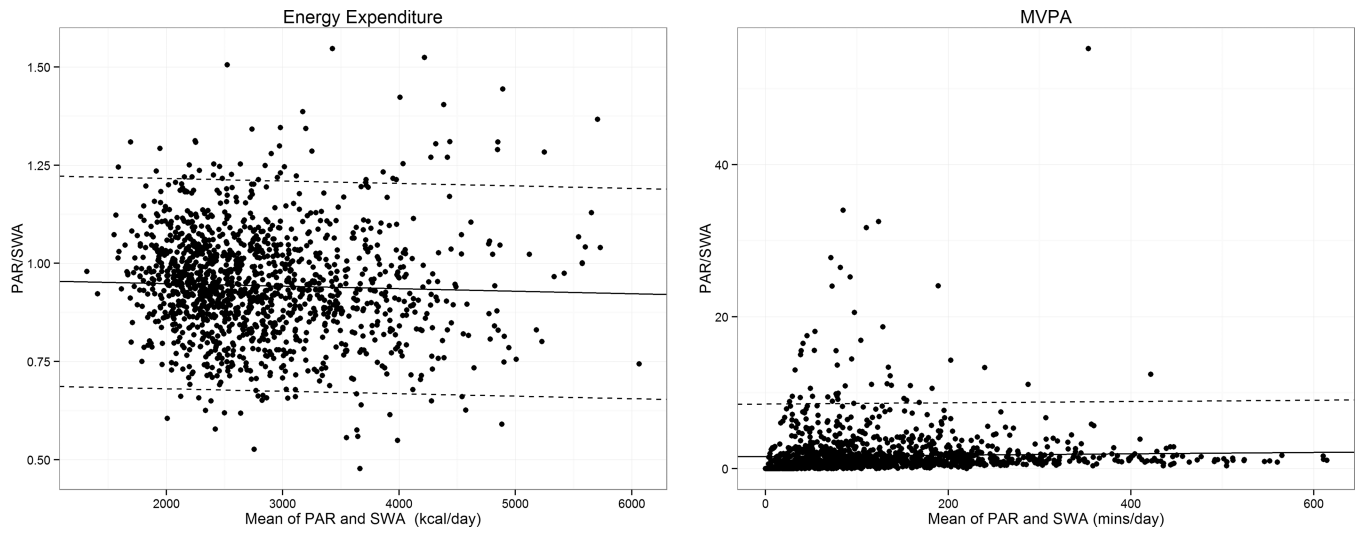
Figure 2.
Bland-Altman plots with 95% limits of agreement between 24-h physical activity recall (PAR) and SenseWear Armband (SWA) for energy expenditure (left) and moderate-to-vigorous physical activity (right).
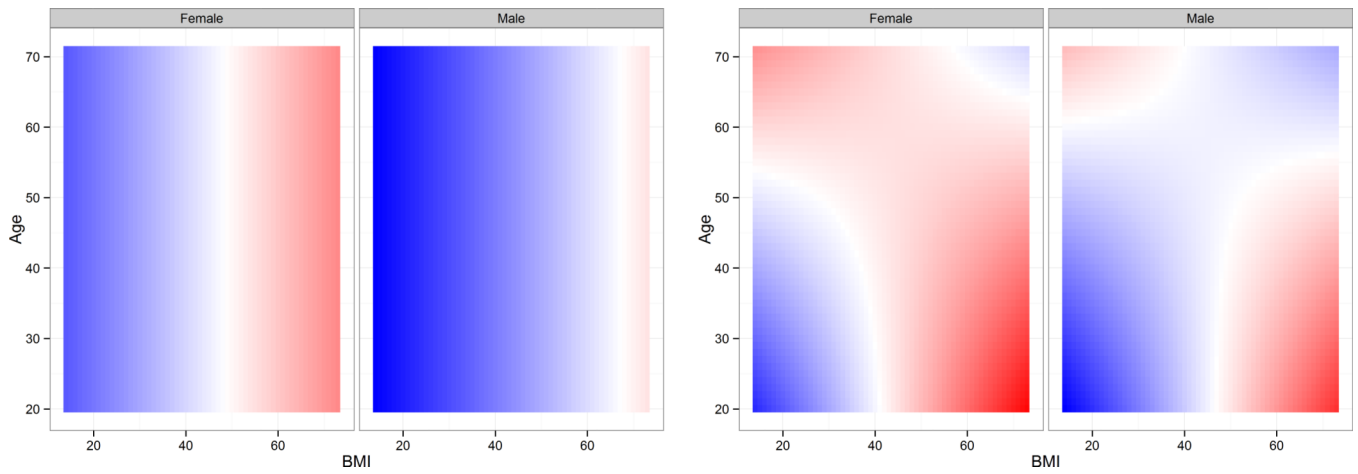
Figure 3.
Heat maps from 3 way ANOVA test – gender × age × body mass index (BMI) – for energy expenditure (top) and moderate-to-vigorous physical activity (bottom). White indicates values from SenseWear Armband (SWA) and 24-hour physical activity recall (PAR) are the same on average; Blue indicates values from PAR are smaller than SWA; Red indicates values from PAR larger than SWA.

Table 1

Characteristics (mean, proportion, and standard error) of the participants included.

Variables	All (n=1347)	Female (n=786; 50.8% (2.2))	Male (n=561; 49.2% (2.2))
Age (yrs)	46.2 (0.3)	47.0 (0.5)	45.4 (0.7)
20–29 yrs,%	8.2 (0.8)	7.0 (1.2)	9.4 (1.3)
30–39 yrs, %	22.0 (1.1)	18.8 (1.9)	25.4 (2.4)
40–49 yrs, %	34.8 (1.1)	38.3 (2.5)	31.1 (2.4)
50–71 yrs, %	35.0 (1.3)	35.9 (2.1)	34.0 (2.8)
Education			
Less than high school, %	3.0 (1.1)	4.0 (1.3)	2.1 (1.0)
High school diploma/some college, %	49.5 (2.0)	50.1 (2.1)	48.9 (3.0)
College/graduate school, %	47.4 (2.0)	46.0 (2.1)	49.0 (3.0)
Ethnicity			
American Indian, %	1.2 (0.6)	0.9 (0.8)	1.6 (1.1)
Asian, %	0.8 (0.3)	0.3 (0.2)	1.4 (0.6)
Black, %	7.2 (1.0)	8.2 (1.6)	6.2 (1.5)
White, %	88.3 (1.2)	88.8 (1.9)	87.7 (1.9)
Others, %	3.2 (0.9)	2.7 (1.1)	3.7 (1.4)
Body Mass Index (BMI)	30.0 (0.2)	29.7 (0.3)	30.2 (0.4)
Normal weight, %	23.2 (1.6)	29.5 (2.6)	19.3 (2.3)
Overweight, %	32.1 (2.0)	30.7 (2.3)	35.3 (3.0)
Obese, %	42.6 (2.0)	39.8 (2.5)	45.4 (3.2)

Note: All values were weighted to account for the complex sampling design. Values in parenthesis represent standard errors unless otherwise indicated.

Table 2

Comparisons between 24-hour physical activity recall and SenseWear Armband by demographic variables for energy expenditure and moderate-to-vigorous physical activity

Group		N	PAR		SWA		ME		MAPE (%) *		Correlations	
			M	SE	M	SE	M	SE	M	SE	Rho	SE
Energy Expenditure (Kilocalories/day)												
All		1347	2821.2	33.8	3049.4	37.5	−228.2	16.0	11.8	0.0	0.83	0.02
Gender	Female	786	2359.6	17.9	2489.5	18.7	−129.9	13.4	10.3	0.0	0.69	0.03
	Male	561	3297.7	57.1	3627.4	47.4	−329.7	26.0	13.3	0.0	0.73	0.03
Age (yrs)	20–29	88	3013.6	128.2	3328.9	118.4	−315.3	87.6	15.0	0.0	0.78	0.06
	30–39	207	3132.0	95.2	3346.5	99.1	−214.5	38.1	11.9	0.0	0.83	0.04
	40–49	305	2804.1	58.7	3051.9	64.1	−247.8	29.9	11.7	0.0	0.80	0.03
	50–71	747	2600.2	36.7	2797.3	44.1	−197.1	16.8	11.1	0.0	0.84	0.02
BMI	14–24	313	2399.7	47.8	2652.3	55.1	−252.6	30.0	12.0	0.0	0.79	0.03
	25–30	432	2759.6	49.4	3006.6	59.4	−247.1	35.8	12.1	0.0	0.80	0.03
	31–73	602	3111.4	55.4	3310.6	63.9	−199.2	28.2	11.4	0.0	0.82	0.02
Moderate-to-Vigorous Physical Activity (Minute/day)												
All		1347	145.8	4.3	131.2	3.5	14.5	3.6	162.1	0.2	0.59	0.03
Gender	Female	786	115.0	5.4	95.6	4.2	19.3	4.1	213.1	0.4	0.47	0.11
	Male	561	177.5	7.6	168.0	4.7	9.5	5.8	109.7	0.2	0.60	0.04
Age (yrs)	20–29	88	181.7	25.0	176.5	16.0	5.1	24.0	108.0	0.2	0.37	0.16
	30–39	207	167.5	15.1	158.7	13.5	8.7	9.4	68.6	0.1	0.71	0.05
	40–49	305	131.7	10.7	127.8	5.8	3.9	9.2	125.1	0.3	0.49	0.07
	50–71	747	137.8	4.5	107.1	4.3	30.7	3.6	269.5	0.7	0.63	0.03
BMI	14–24	313	146.2	9.1	144.8	7.7	1.4	8.7	165.4	0.8	0.57	0.05
	25–30	432	156.1	7.9	136.8	5.42	19.3	6.9	157.8	0.4	0.63	0.04

Group	N	PAR		SWA		ME		MAPE (%) *		Correlations	
		M	SE	M	SE	M	SE	M	SE	Rho	SE
31–73	602	137.3	7.2	119.1	6.93	18.2	6.1	163.6	0.3	0.56	0.05

Note: PAR – 24-hour physical activity recall; SWA – SenseWear Armband; M – mean; SE – standard error; ME – mean error, MAPE; mean absolute percent error.

*
Mean absolute percent error was calculated on a sub-sample of 1337 whose moderate-to-vigorous physical activity minute was not 0 in order to avoid indefinite values.

Table 3

Weighted slopes and intercepts of the fitted regression lines on the Bland–Altman plots for energy expenditure and moderate-to-vigorous physical activity by demographic variables.

Group		Parameter	Energy Expenditure		Moderate-to-Vigorous Physical Activity	
			Regression Coefficients	SE	Regression Coefficients	SE
All		Intercept	0.961	0.021	1.614	0.130
		Slope	-6.3×10^{-6}	7.8×10^{-6}	8.7×10^{-4}	5.9×10^{-4}
Gender	Female	Intercept	1.017	0.038	1.631	0.217
		Slope	-2.3×10^{-5}	1.6×10^{-5}	0.003	0.002
	Male	Intercept	0.817	0.037	1.311	0.239
		Slope	3.0×10^{-5}	1.1×10^{-5}	0.001	7.8×10^{-4}
Age	20–29 yrs	Intercept	0.831	0.066	0.647	0.529
		Slope	2.8×10^{-5}	2.0×10^{-5}	0.005	0.003
	30–39 yrs	Intercept	0.921	0.048	1.285	0.221
		Slope	1.0×10^{-5}	1.5×10^{-5}	5.2×10^{-4}	7.3×10^{-4}
	40–49 yrs	Intercept	0.996	0.040	1.244	0.207
		Slope	-2.0×10^{-5}	1.4×10^{-5}	0.002	0.001
	50–71 yrs	Intercept	1.006	0.032	2.187	0.213
		Slope	-2.2×10^{-5}	1.2×10^{-5}	1.5×10^{-4}	0.001
BMI	14–24	Intercept	1.010	0.056	1.432	0.223
		Slope	-3.5×10^{-5}	2.4×10^{-5}	-5.3×10^{-5}	8.7×10^{-4}
	25–30	Intercept	0.971	0.038	1.754	0.242
		Slope	-1.2×10^{-5}	1.4×10^{-5}	6.4×10^{-4}	5.0×10^{-4}
	31–73	Intercept	0.991	0.038	1.590	0.221
		Slope	-9.8×10^{-6}	1.3×10^{-5}	0.002	0.001

Note: PAR – 24-hour physical activity recall; SWA – SenseWear Armband; M – mean; SE – standard error