Validation of a Computerized 24-Hour Physical Activity Recall (24PAR) Instrument With Pattern-Recognition Activity Monitors

Miguel A. Calabro, Gregory J. Welk, Alicia L. Carriquiry, Sarah M. Nusser, Nicholas K. Beyler, and Charles E. Matthews

Purpose: The purpose of this study was to examine the validity of a computerized 24-hour physical activity recall instrument (24PAR).

Methods: Participants (n = 20) wore 2 pattern-recognition activity monitors (an IDEEA and a SenseWear Pro Armband) for a 24-hour period and then completed the 24PAR the following morning. Participants completed 2 trials, 1 while maintaining a prospective diary of their activities and 1 without a diary. The trials were counterbalanced and completed within a week from each other. Estimates of energy expenditure (EE) and minutes of moderate-to-vigorous physical activity (MVPA) were compared with the criterion measures using 3-way (method by gender by trial) mixed-model ANOVA analyses.

Results: For EE, pairwise correlations were high (r > .88), and there were no differences in estimates across methods. Estimates of MVPA were more variable, but correlations were still in the moderate to high range (r > .57). Average activity levels were significantly higher on the logging trial, but there was no significant difference in the accuracy of self-report on days with and without logging.

Conclusions: The results of this study support the overall utility of the 24PAR for group-level estimates of daily EE and MVPA.

Keywords: motion sensors, physical activity assessment

The development of valid and reliable ways to measure physical activity behavior has been established as an important research objective. Accurate (and cost-effective) measurement techniques would enable researchers to more effectively study the health benefits of physical activity, improve surveillance of physical activity, understand potential correlates of physical activity behavior, and evaluate the effectiveness of behavioral interventions. A variety of methods are available to measure physical activity (PA) and energy expenditure (EE), but each has inherent limitations. Doubly labeled water is widely acknowledged to be the gold standard measure for field-based estimates of EE, but it provides no information about the type, frequency, intensity, and duration of PA. Accelerometers provide a practical method for collecting objective data on physical activity in free-living subjects, but the high cost and inherent challenges associated with processing, reducing, and analyzing the data limit the potential for large-scale surveillance. Although self-report instruments have limitations (outlined later), they remain the most commonly used technique to assess PA in epidemiological and intervention studies. Self-report instruments can be administered to large samples with minor financial costs or burden and also provide valuable information about the context of physical activity that cannot be obtained with other objective methods.

The inherent goal with any measurement instrument is to capture the interindividual variation in behavior and minimize other sources of bias and random error that can affect estimates of parameters that describe the true relationships among study variables. Self-report measures present unique challenges because they are susceptible to a variety of other potential sources of error including systematic reporting of bias and various recall errors. Difficulties in translating frequency and duration information into EE values have also plagued research with self-report instruments.

Researchers in nutrition have experienced many of the same challenges in developing measures of dietary intake. Statistical models have been developed to adjust for self-report errors and bias in dietary intake.
assessments, but to date, these methods have been rarely used in physical activity research. To facilitate the development of measurement error models, it is important to quantify sources of error that may influence physical activity self-report measures. The purpose of the current study was to systematically evaluate the validity of a standardized physical activity self-report measure known as the 24-hour physical activity recall (24PAR) against 2 recently developed pattern-recognition monitors that provide objective information about PA and EE. The 24PAR has been used in a previous epidemiological research study but has not been systematically validated against temporally matched data from objective physical activity monitors. A secondary goal is to examine factors that may influence measurement error in this self-report tool. Comparisons of recall accuracy with and without the use of an activity log are conducted to examine the impact of cueing on recall capability. Comparisons of recalled data from the log trial with data manually entered from the log are conducted to allow us to examine possible sources of recall bias. The insight gained about measurement error in the 24PAR will facilitate the development of comprehensive measurement error models for physical activity self-report.

Methods

Participants

Participants were 20 healthy adult male (n = 10) and female (n = 10) volunteers from a large midwestern university in the United States (age range: 22 to 41 years). The participants included faculty, staff, and graduate students from a diverse array of departments, and efforts were made to recruit participants with varying exercise habits and body sizes. Body mass index (BMI) determined from measured height and weight ranged from 20.0 to 33.0 kg/m² in men (mean = 25.5 ± 4.1) and from 18.0 to 32.0 kg/m² in women (mean = 23.1 ± 4.1). Protocols were approved by the campus Institutional Review Board before beginning the study, and all participants provided written informed consent.

Instruments

Self-Report Measure of Physical Activity—24-Hour Physical Activity Recall (24PAR). The 24PAR is an interviewer-administered 24-hour recall that enables the collection of detailed information about the specific types and intensities of sedentary and active behaviors on the previous day (midnight to midnight). An interviewer systematically leads a participant through a recall of the previous day by prompting them to recall activities performed during 3 distinct time periods (morning, afternoon, evening). For each segment, participants are asked to report time spent sleeping and napping followed by time spent in sedentary behaviors (sitting), standing activities, and various physical activities above 2.5 metabolic equivalents (METs).

Criterion Measures of Physical Activity (IDEEA and SenseWear Pro). Two different monitors were used as criterion measures in the study. The Intelligent Device for Energy Expenditure and Activity (IDEEA) is a portable device that consists of a microcomputer that receives the output signals of 5 small motion sensors attached to the chest, thighs, and feet of the participant. The IDEEA device has been shown to accurately detect (98% accuracy) onset, type, intensity, and duration of PA in laboratory tests and to provide accurate estimates of EE under free-living conditions. The SenseWear Pro2 (SP2) Armband monitor (BodyMedia, Pittsburgh, PA) is a wireless multisensor activity monitor (worn on the upper arm) that relies on accelerometers and several heat-related sensors to detect PA and estimate EE. A number of studies have supported the utility of the SP2 for assessing levels of PA and EE (kcal/min) under free-living conditions. Recent work in our laboratory showed good overall agreement between the SP2 and the IDEEA (average individual minute by minute correlation was r = .76), but the combination of both measures provides for a more comprehensive evaluation.

Procedures

Participants completed 2 trials at least 1 week apart. Participants reported to the laboratory in the morning between 7 and 9 AM on their scheduled day of testing (all trials were performed on weekdays). Anthropometric measures (height and weight) were taken on the participants using research-quality stadiometers and digital scale. The IDEEA and SP2 monitors were then fitted and initialized according to the manufacturers’ recommendations. Participants were told to keep the monitors on until bedtime and to go about their normal daily routine (including some physical activity). Participants were assigned to keep a detailed diary of the activities performed during 1 of the days but not on the other (the order of trials was counterbalanced across the participants to avoid any order effects). In the log, the participants provided the start and end time of each activity, the specific type of activity performed, and the intensity of the activity using a 1 to 4 scale (1 = sitting or standing, 2 = strolling or walking, 3 = brisk walking, 4 = more than level 3).

The subsequent morning, participants reported back to the laboratory to return the monitors and complete a structured 24PAR interview. The 24PAR was administered by a single trained researcher (MC) using a standardized interview protocol established by the

* Internal algorithms provide the estimates of EE from the SP2. Updates are made periodically to improve the accuracy of the algorithms. In the current study we used the latest available version of the Interview Research Software—version 4.1.
The EE estimates (kcal/d) from individual energy expenditure (REE) using the Harris-Benedict for both trials. Four trials had to be repeated because of about the interview. Participants used the same monitors high to very low), and recorded other qualitative notes about the interview. The interviewer also asked about any additional activities that may have been forgotten. The interviewer made an assessment of the overall quality of the information provided (reliable/not reliable) and the level of effort the participant used during the recall (very high to very low), and recorded other qualitative notes about the interview. Participants used the same monitors for both trials. Four trials had to be repeated because of recording problems with the IDEEA monitor.

Data Processing
The data from the computerized 24PAR system were stored in an Access database and were exported to SAS for processing at the completion of all data collection. A customized SAS program developed by a coauthor on the paper (CM) linked activity codes to corresponding MET values from the Compendium of Physical Activity. Because standard MET calculations have been shown to lead to bias, we calculated resting energy expenditure (REE) using the Harris-Benedict equation. The EE estimates (kcal/d) from individual activities were computed as multiples of REE, and total EE was obtained by summing up the EE estimates from all reported activities. To characterize the activity profiles of the sample population, the time devoted to PA during a 24-hour period was summarized using 12 distinct categories (4 types of activity: house—indoor, house—outdoor, occupational/volunteer, and recreation/exercise and 3 intensity levels: light: 1 to 3 METs, moderate: 3 to 6 METs, and vigorous: 6+ METs). The data from the IDEEA and SP2 monitors were saved in individual files and imported into SAS for data processing. Estimates of EE (kcal/d) and time spent in MVPA (min/d) were computed to provide outcome measures that were comparable to those obtained from the 24PAR. Proprietary (internal) algorithms from the activity monitors provided the EE estimates for both the IDEEA and the SP2. The amount of MVPA for each participant was determined by computing the total time (minutes) spent in activities requiring more than 3 METs. Data from the monitors were merged with the 24PAR data in SAS for subsequent analyses.

Data Analysis
Descriptive statistics (sample mean and standard error of the mean) were computed for the primary PA categories to describe the activity profiles of the participants. Descriptive statistics were also computed for overall estimates of EE and MVPA (sample mean and standard error of the mean). Data were also checked for normality to ensure that the distribution of the data would not influence the results.

The primary statistical analyses involved evaluating differences among estimates of EE and MVPA in relation to the 3 PA measurement methods, gender, and the effect of recording PA into a daily log. Mixed-model analyses of variance were used to account for the possible correlation across repeated observations taken on the same individuals in the study. The models (run in SAS 9.0) used participant within gender as person-level random-effect term and the residual variance as a second random-effect term. These analyses assume a common variance for among-person and within-person random effects. The fixed effects included in the models for EE and MVPA were gender, method (24PAR, IDEEA, SP2), and log (participant did or did not record PA in log) as well as the corresponding 2- and 3-way interactions. F tests were used to determine if factors were statistically significant. If a factor was found to be significant, then Tukey-Kramer paired comparisons tests were used to test for differences among levels of fixed effects. Least squares means and standard errors for all effects were estimated under the model and are reported in tables.** Effect sizes are reported to provide a direct indicator of the magnitudes of differences among the different measures of EE and MVPA.

Additional analyses were conducted to determine possible sources of error in the data obtained from the 24PAR. To address the question of whether logging of activity improved accuracy of recall, separate mixed-model analyses were run on the difference scores for EE and MVPA estimates. The fixed effects in these analyses were gender (male or female), log (yes or no), and monitor (SP2 or IDEEA). Correlation analyses were also used to indicate the overall associations among the measures. If difference scores are smaller and correlation coefficients higher on the day that the log was used, it would indicate that the use of a log may be important in improving recall. Direct comparisons were also made between the reported and recalled activity for the same

** Because of the homogeneous variance assumption and balanced sample size among genders and methods, standard errors are the same for all levels in a factor such as PA type and intensity, gender, and method.
day (log condition). Participants turned in their logs but could not use them during the recall. These direct comparisons between the amounts and intensities of activities between the written log and the reported data were conducted to identify possible sources of error in self-report measures of physical activity.

Consistent with contemporary measurement research, we used Bland–Altman graphical procedures to examine agreement across the range of physical activities. In these analyses, the mean of the 2 estimates (x-axis) is plotted against the difference between the 2 estimates (y-axis). This allows systematic forms of bias to be detected. Confidence intervals defining the limits of agreement were established as 1.96 SD from the mean difference.

Results

Participants wore the IDEEA and SP2 monitors for an average of 809 ± 125 minutes during their waking hours (range: 364 to 1006 minutes). Descriptive statistics showing the number of minutes of different types and intensities of activity for men and women are provided in Table 1. The majority of time each day (about 65% of the total monitoring period) was allocated to sitting time. The 3 categories that accounted for the bulk of the activities reported by the participants were light occupational activity, light indoor household activity, and moderate recreational activity. Collectively, these 3 categories accounted for over 70% of all activities reported. ANOVA results indicated that effects due to gender (P = .82), Gender × Type (P = .64), and Gender × Intensity (P = .78) were not significant. This indicates that activity patterns for men and women were similar in this particular sample. Standard tests of normality (Kolmogorov–Smirnov and Shapiro–Wilk) indicated that the data followed a fairly typical normal distribution, so adjustments for skewness were not needed in the subsequent analyses.

Estimates of Energy Expenditure and Physical Activity

Least squares means and standard errors of total EE and MVPA for each treatment condition are provided in Table 2. For EE estimates, there was a significant gender effect, $F_{1,18} = 9.15, P = .007$, with men having higher estimated mean EE levels than women (difference = 597.5 ± 197.5 kcal/d). The method effect and log effects were nonsignificant, indicating that there were no differences across methods or between the 2 conditions in the estimates of EE. There was, however, a significant Gender × Method interaction, $F_{2,90} = 4.06, P = .03$, indicating differential method effects for genders. For men, the SP2 estimate was slightly lower than the 24PAR and the IDEEA, whereas for females, it was slightly higher than the other 2 estimates. No other 2-way (or 3-way) interactions were significant. Bland–Altman plots revealed some tendency for systematic bias across the range of activity levels (Figure 1). The 24PAR tended to underestimate (relative to the IDEEA and SP2) for participants that were less active but to overestimate (relative to the IDEEA and SP2) for participants that were more active. The effect sizes for the pairwise comparisons of EE estimates ranged from -0.14 to 0.17.

For MVPA comparisons, there was a significant method effect, $F_{2,90} = 9.57, P = .002$, and log effect, $F_{1,90} = 8.02, P = .006$, but no gender effects or interactions. Post hoc analyses of the method effect revealed that the

<table>
<thead>
<tr>
<th>Category</th>
<th>Men (n = 10)</th>
<th>Women (n = 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
</tr>
<tr>
<td>Total time reported (h/d)</td>
<td>14.02</td>
<td>0.16</td>
</tr>
<tr>
<td>Sitting reported (h/d)</td>
<td>9.80</td>
<td>0.16</td>
</tr>
<tr>
<td>Physical activity reported (h/d)</td>
<td>3.89</td>
<td>0.16</td>
</tr>
<tr>
<td>House—indoor Light (h/d)</td>
<td>0.97</td>
<td>0.16</td>
</tr>
<tr>
<td>House—indoor Moderate (h/d)</td>
<td>0.24</td>
<td>0.16</td>
</tr>
<tr>
<td>House—indoor Vigorous (h/d)</td>
<td>0.05</td>
<td>0.16</td>
</tr>
<tr>
<td>House—outdoor Light (h/d)</td>
<td>0.35</td>
<td>0.16</td>
</tr>
<tr>
<td>House—outdoor Moderate (h/d)</td>
<td>0.11</td>
<td>0.16</td>
</tr>
<tr>
<td>House—outdoor Vigorous (h/d)</td>
<td>0.05</td>
<td>0.16</td>
</tr>
<tr>
<td>Occupational/Volunteer Light (h/d)</td>
<td>1.46</td>
<td>0.16</td>
</tr>
<tr>
<td>Occupational/Volunteer Moderate (h/d)</td>
<td>0.07</td>
<td>0.16</td>
</tr>
<tr>
<td>Occupational/Volunteer Vigorous (h/d)</td>
<td>0.06</td>
<td>0.16</td>
</tr>
<tr>
<td>Recreation/Exercise Light (h/d)</td>
<td>0.21</td>
<td>0.16</td>
</tr>
<tr>
<td>Recreation/Exercise Moderate (h/d)</td>
<td>1.04</td>
<td>0.16</td>
</tr>
<tr>
<td>Recreation/Exercise Vigorous (h/d)</td>
<td>0.47</td>
<td>0.16</td>
</tr>
</tbody>
</table>
The plot for the SP2 and 24PAR shows a tighter band of error and no evidence of systematic bias (Figure 2 panel B).

To examine agreement in more detail, we also compared the percent of time spent in different intensities of physical activity (rest: <1.5 METs, light: 1.5 to 3.0 METs, moderate: 3.0 to 6.0 METs, and vigorous: >6.0 METs) by the different methods. Separate mixed-model analyses were run to test the agreement between the relative categorization of activity by the 3 methods. We observed a significant Intensity × Method interaction ($P < .001$), but post hoc tests revealed that the only significant differences were between the 24PAR and IDEEA for the rest and light categories. Stacked bar

Only significant pairwise difference for estimates of MVPA was between the IDEEA and SP2 (difference = 40.7 ± 9.3 minutes, $P < .01$). The estimate from the 24PAR fell between the estimates from the pattern-recognition monitor for both trials. The log effect indicates that participants performed more activity when completing a log than when they did not use a log (difference = 21.52 ± 7.60 minutes, $P < .01$). The lack of Method × Log interactions indicates that the effect was evident for each of the measurement methods. The effect sizes for the pairwise comparisons of MVPA estimates ranged from −0.23 to 0.42.

There was wide variability in levels of physical activity across participants, but the estimates from the different methods exhibited similar ranges (24PAR: 5 to 315 minutes, IDEEA: 16 to 182 minutes, SP2: 13 to 275 minutes) and were consistent. Bland–Altman plots for the IDEEA and 24PAR data revealed a tendency for systematic bias, with differences becoming more pronounced at higher levels of activity (Figure 2 panel A). The plot for the SP2 and 24PAR shows a tighter band of error and no evidence of systematic bias (Figure 2 panel B).

### Table 2 Sample Sizes and Least Squares Means ± SE of EE (kcal/d) and MVPA (h/d) for Each Device Overall and Under the Log and No Log Condition Separately

<table>
<thead>
<tr>
<th>Trial</th>
<th>n</th>
<th>Total EE (kcal/d)</th>
<th>Total MVPA (min/d)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>24PAR</td>
<td>IDEEA</td>
</tr>
<tr>
<td>Log condition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>men</td>
<td>10</td>
<td>2307 ± 167.1</td>
<td>2299 ± 167.1</td>
</tr>
<tr>
<td>women</td>
<td>10</td>
<td>1563 ± 167.1</td>
<td>1688 ± 167.1</td>
</tr>
<tr>
<td>both</td>
<td>20</td>
<td>1935 ± 167.1</td>
<td>1993 ± 167.1</td>
</tr>
<tr>
<td>No log condition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>men</td>
<td>10</td>
<td>2190 ± 167.1</td>
<td>2392 ± 167.1</td>
</tr>
<tr>
<td>women</td>
<td>10</td>
<td>1498 ± 167.1</td>
<td>1583 ± 167.1</td>
</tr>
<tr>
<td>both</td>
<td>20</td>
<td>1844 ± 167.1</td>
<td>1987 ± 167.1</td>
</tr>
<tr>
<td>Both conditions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>men</td>
<td>20</td>
<td>2248 ± 149.9</td>
<td>2345 ± 149.9</td>
</tr>
<tr>
<td>women</td>
<td>20</td>
<td>1530 ± 149.9</td>
<td>1635 ± 149.9</td>
</tr>
<tr>
<td>both</td>
<td>40</td>
<td>1889 ± 149.9</td>
<td>1990 ± 149.9</td>
</tr>
</tbody>
</table>

Abbreviations: EE, energy expenditure; MVPA, moderate-to-vigorous physical activity; 24PAR, 24-hour physical activity recall instrument; IDEEA, Intelligent Device for Energy Expenditure and Activity; SP2, SenseWear Pro2.

| a | P < .05 compared with IDEEA. |
| b | P < .05 compared with SP2. |

Only significant pairwise difference for estimates of MVPA was between the IDEEA and SP2 (difference = 40.7 ± 9.3 minutes, $P < .01$). The estimate from the 24PAR fell between the estimates from the pattern-recognition monitor for both trials. The log effect indicates that participants performed more activity when completing a log than when they did not use a log (difference = 21.52 ± 7.60 minutes, $P < .01$). The lack of Method × Log interactions indicates that the effect was evident for each of the measurement methods. The effect sizes for the pairwise comparisons of MVPA estimates ranged from −0.23 to 0.42.

There was wide variability in levels of physical activity across participants, but the estimates from the different methods exhibited similar ranges (24PAR: 5 to 315 minutes, IDEEA: 16 to 182 minutes, SP2: 13 to 275 minutes) and were consistent. Bland–Altman plots for the IDEEA and 24PAR data revealed a tendency for systematic bias, with differences becoming more pronounced at higher levels of activity (Figure 2 panel A). The plot for the SP2 and 24PAR shows a tighter band of error and no evidence of systematic bias (Figure 2 panel B).

To examine agreement in more detail, we also compared the percent of time spent in different intensities of physical activity (rest: <1.5 METs, light: 1.5 to 3.0 METs, moderate: 3.0 to 6.0 METs, and vigorous: >6.0 METs) by the different methods. Separate mixed-model analyses were run to test the agreement between the relative categorization of activity by the 3 methods. We observed a significant Intensity × Method interaction ($P < .001$), but post hoc tests revealed that the only significant differences were between the 24PAR and IDEEA for the rest and light categories. Stacked bar
model analyses as well as with correlation analyses. The mixed-model analyses tested whether difference scores between estimates of EE and MVPA were significantly different between the log and no log conditions. For both comparisons, we found the main effects for the log effect to be nonsignificant. Correlation analyses confirm that there was similar agreement between the measures for both the log and no log condition (see Table 3). The correlations between the 24PAR and the SP2 monitor were slightly higher. Correlations with the IDEEA were somewhat lower and more variable for estimates of MVPA. Overall, the correlations between the 24PAR and the SP2 monitor were slightly higher.

plots of the average percent of individual activity for each level reveal that the IDEEA data resulted in lower amounts of rest but higher amounts of light activity relative to the 24PAR and SP2 (see Figure 3). There were no differences between the allocations from the 24PAR and the SP2 for any of the intensities and no pairwise differences among any method for the amount of moderate or vigorous activity. These effects were similar for both men and women and were consistent across the 2 trials.

Effect of Logging on Accuracy of Self-Reporting

The potential effect of logging on improving accuracy of self-reporting was tested with 2 separate mixed-
Figure 2 — Bland–Altman Plots for total physical activity estimates. (A) Comparisons between 24PAR and IDEEA. (B) Comparisons between 24PAR and SP2.

Figure 3 — Distribution of time spent in different categories of physical activity for the 24PAR, IDEEA, and SP2.
The agreement in 2 different trials (1 trial involved IDEEA and PA. Another unique feature is that we examined the amount of rest. The extra amount of light activity estimated the amount of light activity and underestimated the monitors. The 24PAR estimates were consistently lower than the IDEEA estimates, but the pattern was not consistent for comparisons with the SP2 (the 24PAR tended to underestimate relative to the SP2 for men and overestimate relative to the SP2 for women). Because this effect was not noted in comparisons with the IDEEA (or for the MVPA analyses), it is likely that the results are due to variability in SP2 estimates for men and women rather than to measurement error in the 24PAR.

Consistent with most activity-assessment research, we observed considerable individual variability in EE comparisons allowed us to evaluate whether these factors influence the accuracy of self-report measures of PA.

The 24PAR was found to be highly correlated with both the SP2 and the IDEEA for estimates of EE. The absolute differences in group-level EE estimates were relatively small (about 100 kcal/d or 5% to 6% error), and these differences were not statistically significant. The reported effect sizes for the pairwise comparisons were less than 0.17, indicating that these differences are small in both real terms and practical terms.

The results revealed a significant Gender × Method interaction, suggesting that there were differential gender effects for estimates of EE from the 24PAR or the monitors. The 24PAR estimates were consistently lower than the IDEEA estimates, but the pattern was not consistent for comparisons with the SP2 (the 24PAR tended to underestimate relative to the SP2 for men and overestimate relative to the SP2 for women). Because this effect was not noted in comparisons with the IDEEA (or for the MVPA analyses), it is likely that the results are due to variability in SP2 estimates for men and women rather than to measurement error in the 24PAR.

Consistent with most activity-assessment research, we observed considerable individual variability in EE estimates across the methods. The differences in individual EE estimates from the IDEEA and the 24PAR ranged from 686.0 kcal/d to −545.9 kcal/d. The differences in individual EE estimates from the 24PAR and the SP2 ranged from 945.9 kcal/d to −662.5 kcal/d. This confirms that there are large individual differences in estimates from these different assessments—despite small group-level differences. Overall measurement agreement among methods was somewhat weaker for estimates of MVPA, and there was also considerable individual variability in these estimates.

It is noteworthy that the mean 24PAR values were between the estimate from the IDEEA and the SP2 for both men and women. In fact, there were no significant differences between the 24PAR and the 2 pattern-recognition monitors, but the 2 monitors were found to yield significantly different estimates. To examine the agreement with the 24PAR in more detail, we compared the distribution of activity into the categories of rest, light, moderate, and vigorous. The stacked bar graphs, correlations, and Bland–Altman plots showed that there was good overall agreement between the SP2 and the 24PAR for different intensities of physical activity but worse correspondence between the 24PAR and the IDEEA. Compared with the SP2, the IDEEA overestimated the amount of light activity and underestimated the amount of rest. The extra amount of light activity explains why the IDEEA led to higher estimates of EE compared with the other measures. Overall, the 24PAR showed better agreement with the SP2 than the IDEEA for estimating MVPA (see Figure 3).

Collectively, the results extend previous studies that have demonstrated the utility of the 24PAR. The reported

### Table 3 Correlations Between Pattern-Recognition Monitors and 24PAR for Total EE (kcal/d) and MVPA (min)

<table>
<thead>
<tr>
<th>Trial</th>
<th>Total EE (kcal · kg⁻¹ · d⁻¹)</th>
<th>IDEEA</th>
<th>SP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log</td>
<td>20</td>
<td>.91b</td>
<td>.90b</td>
</tr>
<tr>
<td>No log</td>
<td>20</td>
<td>.89b</td>
<td>.89b</td>
</tr>
<tr>
<td>All</td>
<td>40</td>
<td>.89b</td>
<td>.89b</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trial</th>
<th>MVPA (min)</th>
<th>IDEEA</th>
<th>SP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log</td>
<td>20</td>
<td>.67b</td>
<td>.70b</td>
</tr>
<tr>
<td>No log</td>
<td>20</td>
<td>.48a</td>
<td>.61b</td>
</tr>
<tr>
<td>All</td>
<td>40</td>
<td>.57a</td>
<td>.66a</td>
</tr>
</tbody>
</table>

Abbreviations: EE, energy expenditure; MVPA, moderate-to-vigorous physical activity; 24PAR, 24-hour physical activity recall instrument; IDEEA, Intelligent Device for Energy Expenditure and Activity; SP2, SenseWear Pro2.

a Significant at P < .05.
b Significant at P < .01.

(average r = .66) than the corresponding values for the IDEEA (average r = .57).

Additional analyses were conducted to compare the type and amount of activity reported on the log versus the activities that were recalled during the interview. There were no significant differences between the recalled/recorded amount of light (P = .54), moderate (P = .98), or vigorous activity (P = .74) or the amount of sitting time (P = .47) during the day. Participants did report slightly more walking on the day that they recorded on the log (P = .05), but the magnitude of the difference was small (~12 minutes). When examining types of activities, significant differences were found for the recalled/recorded amount of light household activity (P = .03) and light occupational activity (P = .04) performed during the day. In both cases, participants tended to recall slightly higher amounts of light activity than they recorded on the log. The errors in the recalled/recorded activity patterns appear to be the result, in part, of differences in the total amount of time that was recorded. Participants had an average of 35 minutes of unrecorded activity on their logs, and this was likely allocated into light household activity or light occupational activity when they were requested to provide a complete report for the activity they performed during the interview.

### Discussion

The primary purpose of this study was to evaluate the validity of the 24PAR for estimating EE and MVPA. A unique aspect of the study is that comparisons were made against 2 different pattern-recognition devices that have been shown to provide accurate estimates of EE and PA. Another unique feature is that we examined the agreement in 2 different trials (1 trial involved recording activity in a log and the other did not). These comparisons allowed us to evaluate whether these factors influence the accuracy of self-report measures of PA.

The 24PAR was found to be highly correlated with both the SP2 and the IDEEA for estimates of EE. The absolute differences in group-level EE estimates were relatively small (about 100 kcal/d or 5% to 6% error), and these differences were not statistically significant. The reported effect sizes for the pairwise comparisons were less than 0.17, indicating that these differences are small in both real terms and practical terms.

The results revealed a significant Gender × Method interaction, suggesting that there were differential gender effects for estimates of EE from the 24PAR or the monitors. The 24PAR estimates were consistently lower than the IDEEA estimates, but the pattern was not consistent for comparisons with the SP2 (the 24PAR tended to underestimate relative to the SP2 for men and overestimate relative to the SP2 for women). Because this effect was not noted in comparisons with the IDEEA (or for the MVPA analyses), it is likely that the results are due to variability in SP2 estimates for men and women rather than to measurement error in the 24PAR.

Consistent with most activity-assessment research, we observed considerable individual variability in EE comparisons allowed us to evaluate whether these factors influence the accuracy of self-report measures of PA.

The 24PAR was found to be highly correlated with both the SP2 and the IDEEA for estimates of EE. The absolute differences in group-level EE estimates were relatively small (about 100 kcal/d or 5% to 6% error), and these differences were not statistically significant. The reported effect sizes for the pairwise comparisons were less than 0.17, indicating that these differences are small in both real terms and practical terms.

The results revealed a significant Gender × Method interaction, suggesting that there were differential gender effects for estimates of EE from the 24PAR or the monitors. The 24PAR estimates were consistently lower than the IDEEA estimates, but the pattern was not consistent for comparisons with the SP2 (the 24PAR tended to underestimate relative to the SP2 for men and overestimate relative to the SP2 for women). Because this effect was not noted in comparisons with the IDEEA (or for the MVPA analyses), it is likely that the results are due to variability in SP2 estimates for men and women rather than to measurement error in the 24PAR.

Consistent with most activity-assessment research, we observed considerable individual variability in EE estimates across the methods. The differences in individual EE estimates from the IDEEA and the 24PAR ranged from 686.0 kcal/d to −545.9 kcal/d. The differences in individual EE estimates from the 24PAR and the SP2 ranged from 945.9 kcal/d to −662.5 kcal/d. This confirms that there are large individual differences in estimates from these different assessments—despite small group-level differences. Overall measurement agreement among methods was somewhat weaker for estimates of MVPA, and there was also considerable individual variability in these estimates.

It is noteworthy that the mean 24PAR values were between the estimate from the IDEEA and the SP2 for both men and women. In fact, there were no significant differences between the 24PAR and the 2 pattern-recognition monitors, but the 2 monitors were found to yield significantly different estimates. To examine the agreement with the 24PAR in more detail, we compared the distribution of activity into the categories of rest, light, moderate, and vigorous. The stacked bar graphs, correlations, and Bland–Altman plots showed that there was good overall agreement between the SP2 and the 24PAR for different intensities of physical activity but worse correspondence between the 24PAR and the IDEEA. Compared with the SP2, the IDEEA overestimated the amount of light activity and underestimated the amount of rest. The extra amount of light activity explains why the IDEEA led to higher estimates of EE compared with the other measures. Overall, the 24PAR showed better agreement with the SP2 than the IDEEA for estimating MVPA (see Figure 3).

Collectively, the results extend previous studies that have demonstrated the utility of the 24PAR. The reported
validity coefficients were higher than previously reported by Matthews and colleagues\textsuperscript{24} for comparisons between the 24PAR instrument and accelerometers (CSA, Actillume). In this study, low to moderate correlations were reported for both total activity ($r = .31$) and sitting time ($r = .47$).\textsuperscript{24} Values in the current study were moderate for MVPA and high for estimates of EE and sitting time. The lack of significant differences for estimates of EE and MVPA (relative to the IDEEA and SP2) also provide good support for the overall utility of the 24PAR. The use of 2 criterion measures (IDEEA and SP2) strengthens the results because we demonstrated that the 24PAR agrees reasonably well with both comparison measures—despite some differences in agreement between the 2 criterion measures.

Although there were some subtle gender differences and discrepancies highlighted in the results, the results are impressive considering the challenges of capturing detailed information about activity and energy expenditure in free-living populations. The agreement is especially impressive considering that the 24PAR and pattern-recognition devices arrive at their estimates in different ways. The MVPA estimates from the 24PAR, for example, are based directly on the reported information, whereas the MVPA estimates from the monitors are obtained by extrapolation from the minute-by-minute EE estimations. The EE estimates from the 24PAR, in turn, rely on MET estimates from the Compendium of Physical Activity, whereas the monitors estimate EE based on pattern-recognition algorithms. Although there was considerable individual variability in estimates of both EE and MVPA, much of the variability may be the result of inherent limitations in the criterion measures as well as to limitations of the 24PAR coding procedures and MET-based estimations. The IDEEA, in particular, appears to overestimate the amount of light activity—presumably by overestimating the energy cost of some rest activities. This likely led to some of the reported differences between the amounts of rest and light activity and explains some of the variability in overall estimates of EE among methods.

A secondary goal of the study was to advance understanding of factors that influence error in self-report instruments such as the 24PAR. The comparisons of agreement on trials with and without a log were designed to determine if recall would be enhanced by tracking activity with a log. We observed moderate to high correlations between the methods for both the log and no-log conditions. There was also no significant difference in the accuracy of EE and MVPA estimates from the 24PAR between the 2 conditions. We did observe a significant difference in the amount of activity performed on the log and no-log condition (a difference of about 20 minutes), indicating that participants may have been more inclined to be active if using a log. There was no log effect for the estimates of EE, but this may be due to the fact that rest and light activity make larger contributions to the total estimates of EE than moderate or vigorous activity—thereby obscuring the difference in activity that was performed. The Bland–Altman plots also provided information about the impact of logging. We observed no difference in the distribution of data points on the Bland–Altman plots between the conditions. Collectively, these results suggest that logging did not improve accuracy of recall.

Direct evaluation of recall accuracy was obtained by comparing a participant’s reported activity on the log with the activity that they recalled on the 24PAR. There were no significant differences in reported amounts of rest, light, moderate, and vigorous activity but some differences in the reported categorization of PA. Participants tended to overestimate the amount of light occupational activity and light household activity compared with what was reported on the log. The differences are due, in part, to the lack of complete data on the log but may also reflect the cognitive challenges associated with recalling information about common activities of daily living that may not have been consciously attended to while being performed. Studies have frequently noted that recall accuracy is higher for vigorous activity than moderate activity, so it is not surprising that there are some discrepancies for the reported amounts of light activity. However, the good overall correspondence between the participant’s recall of their activity and a manual coding of their objective activity logs provides evidence that participants were able to provide appropriate details about activity patterns on the previous day. The results are consistent with findings from Timperio and colleagues,\textsuperscript{27} who concluded that the use of a logbook does not significantly influence results from recall-based activity assessments. However, in contrast with this study, we found that logging may prompt an increase in physical activity. Because there was no Method × Log interaction, the difference cannot be attributed to social desirability in the self-reported activity following logging.

Overall, the 24PAR appears to be a promising tool for assessing MVPA and EE in population-based research studies. The primary source of variability in the data appears to be the coding algorithms used to convert between activity data and energy expenditure data. The 24PAR relies on rough estimates from the Compendium of Physical Activities to convert estimates into EE. The criterion measures, in turn, use internal algorithms to convert movement into estimates of EE and corresponding estimates of MVPA. A key limitation in the study is the narrow characteristics of the sample population. Most participants were college-educated individuals, and it is possible that this population would be more compliant or more reflective when faced with a complex recall task. Because of the restricted nature of our sample, it is possible that the results would not generalize to populations with more diverse characteristics. Future studies should use a more diverse population to evaluate the utility for different participant characteristics (ie, levels of education, occupations, and levels of PA).
References


