Reliability of Accelerometry-Based Activity Monitors: A Generalizability Study

GREGORY J. WELK[1], JODEE A. SCHABEN[1], and JAMES R. MORROW, JR.[2]

[1]*Iowa State University, Ames, IA; and* [2]*University of North Texas, Denton, TX*

ABSTRACT

WELK, G. J., J. A. SCHABEN, and J. R. MORROW, JR. Reliability of Accelerometry-Based Activity Monitors: A Generalizability Study. *Med. Sci. Sports Exerc.*, Vol. 36, No. 9, pp. 1637–1645, 2004. **Introduction:** Numerous studies have examined the validity of accelerometry-based activity monitors but few studies have systematically studied the reliability of different accelerometer units for assessing a standardized bout of physical activity. Improving understanding of error in these devices is an important research objective because they are increasingly being used in large surveillance studies and intervention trials that require the use of multiple units over time. **Methods:** Four samples of college-aged participants were recruited to collect reliability data on four different accelerometer types (CSA/MTI, Biotrainer Pro, Tritrac-R3D, and Actical). The participants completed three trials of treadmill walking (3 mph) while wearing multiple units of a specific monitor type. For each trial, the participant completed a series of 5-min bouts of walking (one for each monitoring unit) with 1-min of standing rest between each bout. Generalizability (G) theory was used to quantify variance components associated with individual monitor units, trials, and subjects as well as interactions between these terms. **Results:** The overall G coefficients range from 0.43 to 0.64 for the four monitor types. Corresponding intraclass correlation coefficients (ICC) ranged from 0.62 to 0.80. The CSA/MTI was found to have the least variability across monitor units and trials and the highest overall reliability. The Actical was found to have the poorest reliability. **Conclusion:** The CSA/MTI appeared to have acceptable reliability for most research applications (G values above 0.60 and ICC values above 0.80), but values with the other devices indicate some possible concerns with reliability. Additional work is needed to better understand factors contributing to variability in accelerometry data and to determine appropriate calibration protocols to improve reliability of these measures for different research applications. **Key Words:** PHYSICAL ACTIVITY ASSESSMENT, ACCELEROMETER, CSA/MTI, BIOTRAINER, TRITRAC, ACTICAL

Accelerometry-based activity monitors are one of the most commonly used methods for assessing free-living physical activity (19). Depending on the device and the recording interval, they can store continuous data for up to 30 d and be downloaded to a computer for data processing. Contemporary units are fairly small and are generally well tolerated by participants, so accelerometers have become an accepted means of capturing objective data on physical activity behavior.

Considerable research has been conducted to examine the validity of accelerometry-based activity monitors. Validity has generally been found to be strong under laboratory conditions (8,16,18,20), but a number of studies have described the inability of accelerometers to accurately assess "lifestyle" activities that involve large upper-body activity (1,10,21). Overall, the general consensus in the literature is that monitors can provide an accurate indicator of energy expenditure for locomotor activities but less precise estimates for free-living activity (19). Comparatively, much less is known about the reliability of accelerometry-based devices. Understanding sources of error and reliability of these devices is essential because this ultimately sets the limit on validity. Some may assume that reliability must be good if validity is acceptable, but this cannot be assumed because most validity studies have utilized a single unit (or several units) to test validity and/or to establish calibration equations. As research begins to employ accelerometers in surveillance research and intervention research, it becomes increasingly important to understand how much variability exists between units and how variable responses may be over time.

The most common approach in reliability studies has been to use a mechanical setup that allows for a standardized amount of movement. Kochersberger et al. (12) reported intraclass correlation coefficients of $R = 0.97$ when comparing output from nine different Tritrac units using a mechanical shaker table. In a similar design, Nichols et al. (16) reported nonsignificant differences in accelerometer output and a coefficient of variation of 1.79% when comparing four Tritrac units. Metcalf et al. (14) evaluated the reliability of the original Computer Science Application (CSA) monitor (now also referred to as the MTI monitor) using a turntable device. They reported low intra-instrument coefficients of variation (1.83%) and slightly higher inter-instrument coefficients of variation of ~5%. Brage et al. (4) used an oscillating mechanical device to examine the reliability of the CSA/MTI monitor across a range of different simulated

Address for correspondence: Gregory J. Welk, Ph.D., Department of Health and Human Performance, Iowa State University, 257 Forker Building, Ames, IA 50011; E-mail: gwelk@iastate.edu.
Submitted for publication February 2004.
Accepted for publication April 2004.

0195-9131/04/3609-1637
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/01.MSS.0000074670.03001.98

vertical accelerations. They confirmed that variability between CSA/MTI units is greater than variability within units and reported that each unit was significantly different from the mean value from all of the units. They also noted considerable heteroscedasticity across units in the response to different acceleration values and recommended unit specific calibrations as a way to obtain more accurate data for field research.

Several research groups have examined reliability (objectivity) by comparing outputs from units worn on opposite hips. Nichols et al. (16) reported intraclass correlation coefficients (ICC) ranging from R = 0.73 to R = 0.87 for two different Tritrac monitors worn during free living activity. Trost et al. (18) reported average ICC values of 0.87 for the CSA/MTI in a calibration study with youth. Welk et al. (20) reported average ICC values of R = 0.66 and R = 0.84 for the Actitrac and Biotrainer monitors in a laboratory calibration study. In this study, significant differences were reported in the actual accelerometer counts from monitors worn on the right and left side but the effect sizes were small. Measurement experts generally suggest that reliability be above 0.80 but the precision needed may depend on the research application. Based on these criteria, these side-to-side comparisons suggest that monitors may be reasonably reliable but Jakicic et al. (11) reported that reliability is likely to be higher for walking and running (range: R = 0.76–0.92) than for the stepping, sliding, or cycling tasks (range: R = 0.54–0.88). Another limitation of these types of comparisons is that it is only possible to compare two units at a time. It is also not clear how variable responses are if similar data were obtained on multiple occasions. To continue to advance research with accelerometry-based activity monitors, more systematic studies of reliability are needed.

Direct comparisons between different types of commercially available monitors are also needed. A number of researchers have sought to determine whether three-dimensional monitors (e.g., Tritrac or Tracmor) may offer advantages in assessing physical activity compared to unidimensional devices like the CSA/MTI and Biotrainer, but the results have been equivocal. Some studies have indicated advantages for the Tritrac (2,5,6) whereas other studies (13,22) have demonstrated high correlations (r > 0.88) between unidimensional devices and the three-dimensional Tritrac monitor, indicating that they provide similar information. A possibility that has not been examined is whether the three-dimensional signal may help to provide a more reliable indicator of activity compared with uni-axial monitors. This is a tenable hypothesis as the three-dimensional output variable from the Tritrac (vector magnitude) should theoretically be independent of the orientation of the accelerometer on the body. Direct comparisons between different types of monitors are needed to compare reliability of these devices.

The purpose of this study was to systematically examine the reliability of four commercially available activity monitors (Tritrac, CSA/MTI, Biotrainer, and Actical) for a structured bout of physical activity. Data were obtained from 7 to 10 units of each monitor type and comparisons were made over multiple trials to examine the reproducibility of the results. This allowed us to examine interunit variability (i.e., differences between units of a given monitor) as well as intra-individual variability (differences between individual responses to a given monitor). Generalizability theory was used to take advantage of the fully crossed design (i.e., multiple monitors worn on repeated trials). This measurement approach allows the various sources of variability in accelerometer data to be partitioned and quantified (9).

METHODS

Participants

Participants were college students majoring in Exercise Science from a large Midwestern university. The project was conducted over a span of 2 yr with different samples of participants evaluating a different monitor each semester. Because the focus of the study is on reliability, the comparisons of interest are based on repeatability of the output for an individual's own data. Therefore, the use of separate samples is not a major limitation in the study. The samples ranged from 32 to 38 participants. Each sample had a higher percentage of females (range: 67% to 78%) due to the demographics of the population used for recruitment (an undergraduate exercise science class), and the greater interest in females in volunteering for the project. All participants completed written informed consent documents and the overall study protocol was approved by the Institutional Review Board at the lead author's institution.

The descriptive characteristics of the study populations are shown in Table 1. A two way (gender × sample) MANOVA was performed on the anthropometric variables in the Table to check for any systematic differences in the sample populations. The overall multivariate test revealed no significant gender × sample interactions ($P = 0.499$) and no significant sample main effects ($P = 0.924$). There was a significant gender main effect [$F(3,127) = 53.98$)] with males being significantly taller, heavier, and having higher BMI values than females. This was true for each of the four samples.

Instruments

CSA/MTI (Actigraph): sample 1. The CSA/MTI monitor (Manufacturing Technology, Inc., Fort Walton Beach, FL) is a one-dimensional accelerometer that uses an infrared interface for data transfer. The small size (2 × 1.5 × 0.6 inches) and durable case has made this monitor popular in behavioral research. The unit allows users to specify the recording epochs (from 5 s to 1 min) and to program the start time during the initialization phase. The most commonly used version of the CSA/MTI (Model 7164 Actigraph) has 64kbyte of memory and can record continuous data for 22 d using 1-min epochs. The validity of the CSA/MTI has been evaluated in a number of studies (8,13,17,18,21), but less work has been done to examine

TABLE 1. Characteristics of sample populations used for each monitor.

Monitor	Sample 1 CSA/MTI Mean (SD)	Sample 2 Biotrainer Mean (SD)	Sample 3 Actical Mean (SD)	Sample 4 Tritrac Mean (SD)
Females	*N* = 25	*N* = 23	*N* = 26	*N* = 22
Height (m)	1.69 (0.07)	1.68 (0.06)	1.68 (0.07)	1.69 (0.07)
Weight (kg)	63.51 (9.19)	62.39 (9.23)	65.22 (9.06)	64.44 (7.38)
BMI	22.10 (2.17)	22.14 (3.46)	23.06 (3.22)	22.58 (2.82)
Males	*N* = 7	*N* = 11	*N* = 12	*N* = 11
Height (m)	1.79 (0.06)	1.80 (0.05)	1.81 (0.06)	1.81 (0.06)
Weight (kg)	92.99 (12.72)	86.03 (12.51)	88.94 (16.26)	87.73 (17.21)
BMI	29.11 (4.11)	26.65 (3.68)	26.90 (3.51)	26.40 (3.60)
Combined	*N* = 32	*N* = 34	*N* = 38	*N* = 33
Height (m)	1.71 (0.08)	1.72 (0.08)	1.72 (0.09)	1.73 (0.09)
Weight (kg)	69.96 (15.81)	70.04 (15.29)	72.72 (16.08)	72.20 (15.89)
BMI	23.63 (3.95)	23.59 (4.08)	24.27 (3.73)	23.85 (3.54)

reliability. Ten different CSA/MTI monitors were compared in the study.

Biotrainer Pro: sample 2. The Biotrainer Pro (IM Systems, Baltimore, MD) is a uniaxial device that is larger than the CSA/MTI monitor but approximately half the cost. It is considered to be a uniaxial device but the accelerometer is actually positioned 45° to vertical in the sagittal plane to pick up two dimensions of movement. The device is downloaded to a computer using a serial port connection with the device. The Biotrainer Pro monitors can be programmed to record in epochs ranging from 15 s to 5 min and the sensitivity can be changed depending on the type of activity to be performed. The device can store data for up to 22 d using 1-min epochs. Evidence of convergent validity was reported in a previous study by the lead author (21). This study reported high correlations between the Biotrainer Pro and both the CSA/MTI and Tritrac under treadmill and field conditions (range of correlations: r = 0.74–0.85). A subsequent study (20) reported strong validation criteria (R^2 = 0.88) and cross validation results (R^2 = 0.86) for a treadmill based calibration equation for the Biotrainer. No information is currently available on the reliability of this device. Nine different Biotrainer Pro monitors were compared in the study.

Tritrac: sample 3. The Tritrac R3D is a three-dimensional monitor that was based on the same principles as the original Caltrac monitor. The Tritrac provides accelerometry counts for all three dimensions (side-to-side: X; forward and backward: Y; and up and down: Z) and a composite indicator of movement known as vector magnitude. The device downloads to a computer using a parallel port interface. The Tritrac R3D can be programmed to record in epochs ranging from 1 s to 1 min and can store up to 7 d of data at the 1 min setting. A study by Nichols et al. (16) provided the most frequently used calibration equation for the Tritrac, and several studies have reported good results with this device (5,13,21). A newer version of the Tritrac called the RT3 has been released by StayHealthy, Inc. (Monrovia, CA). This version is considerably smaller than the R3D and employs a single triaxial accelerometer instead of three individual accelerometers soldered together. At present, the authors are not aware of any studies that have reported the reliability or validity of the new RT3 units. The original Tritrac R3D is still widely used by many research groups, so comparisons were made in the present study with nine different Tritrac R3D units.

Actical: sample 4. The Actical (Mini-Mitter, Sunriver, OR) is considered to be an "omni-directional" device capable of recording in all directions. It is currently the smallest, commercially available, accelerometry-based activity monitor (the dimensions are 28 mm × 27 mm × 10 mm, and it weighs 17.5 g). The device downloads to a computer using a telemetry-based receiver that connects to a computer using a serial port connection. The device can be programmed to record in epochs ranging from seconds to min and stores up to 44 d of data at a 1-min epoch. The Actical has not been widely used but one study (7) reported good agreement between the Actical counts and a direct observation measure in children. Nine Actical units were originally used in the original design but two of the devices did not get downloaded properly in the course of the study and data were lost. To provide a more representative sample, we used the available data from the remaining seven monitors rather than the smaller sample of participants that had complete data with nine instruments.

Data Collection Procedures

The participants in the study completed three trials of treadmill walking at 3 mph while wearing one of the four types of accelerometers on the right hip. The individual trials were divided into a series of 5-min bouts of walking, with 1 min of standing rest between each bout. Participants wore different units of a specific monitor type for each 5 min bout and then straddled the treadmill to switch units during the 1-min resting phase. The individual units were attached to a waist-worn belt using clips or straps to facilitate quick changes. Positioning of all monitor units was standardized at the mid-axillary line for each monitor type and for all three trials with a given monitor. Three different treadmills were used in the study to facilitate data collection, but each participant completed all three trials on the same treadmill to avoid additional sources of variability in the results. During each trial, the speed of the treadmill was checked with a timing device to ensure similar conditions for each trial. The order in which the monitors were worn was also varied across participants to avoid any order effect related to fatigue or changing body mechanics. Although

data were collected in different semesters, the protocols used to collect data on the four different monitors were identical.

Data Analysis

Data from each monitor were downloaded to a computer and imported into Microsoft Excel for initial data processing. Data from each bout of activity were identified by time and values for the 5 min of walking were transferred to a different data set for subsequent processing. Temporal processing of the data in this manner was possible because the monitors were always initialized to record data in 60-s epochs, and each trial was started at the beginning of a new minute. To avoid any influence due to starting or stopping on the treadmill, the average counts from minutes 2 to 4 were used to represent the activity count for each bout. A two-way (treadmill × gender) ANOVA was performed with the data from each monitor type to examine potential differences in values recorded on the three different treadmills and to test for differential responses between males and females.

Descriptive statistics (mean and SD) were computed separately for each unit (of a given monitor type) and for each trial. Coefficients of variation (CV = SD/mean) were used to examine the variability across monitors as well as within and between individuals. Because the CV is independent of the units of measurement, it provides the most effective descriptive statistic to summarize variability. A CV was computed across the multiple units to examine the variability among the units for a given monitor type. This calculation utilized the participants' mean accelerometer count across the three trials in order to have the most stable indicator and to avoid using multiple trials in the calculation. A CV value was computed across the different units worn by each participant and the mean CV across participants was reported for each trial. The mean CV across individuals for a given monitor provides an indication of the variability due to subject differences. Correlations between accelerometer counts and anthropometric variables were computed to help explain variability in accelerometer counts for a standardized bout of activity.

Generalizability (G), an extension of intraclass reliability, was used to more quantitatively partition the total variability associated with the accelerometer data. Analyses were performed separately for each monitor using a two-way (monitor by trial) ANOVA design. Monitors and trials were considered random facets in the fully crossed design. Expected mean squares and variance components were estimated based on procedures described by Morrow (15). The G-study phase of the work resulted in percents of variance associated with each facet and interaction in the model. The obtained G coefficients are interpreted like reliability coefficients with perfect generalizability being 1.0 and no generalizability being 0.0. Based on the G coefficients and facets identified as important contributions to model error obtained in the G-study, a D-study phase is conducted. The D-study phase of the generalizability analysis can be used to determine the generalizability (i.e., reliability) when the number of monitors or trials is changed. Increasing or decreasing the number of facet levels helps identify the minimal and/or optimal number of facet levels required to establish the desired generalizability. In this study, the D-study component provides the generalizability estimate for a single trial and a single monitor because that is how these devices are commonly used in practice.

RESULTS

Descriptive results. Statistical analyses were restricted to only those participants with complete data on three different trials. The samples used to test each monitor varied in size but were similar in gender composition and anthropometric characteristics (See Table 1).

Separate two-way (gender × treadmill) ANOVA were performed to check for any systematic variability due to the use of different treadmills and any gender effects. The separate multivariate tests were significant for the Tritrac [$F(5,32) = 3.03$, $P = 0.027$] and approached significance for the Actical [$F(5,37) = 2.29$, $P = 0.069$]. For both monitors, there was a significant main effect for gender with females having significantly higher ($P < 0.05$) activity counts than the males (Tritrac: effect size = 0.81; Actical: effect Size = 0.60). The effect sizes (ES) for the two monitors with nonsignificant gender comparisons were low (ES = −0.22 for the CSA/MTI and ES = 0.10 for the Biotrainer). The treadmill main effect was not significant for any of the comparisons indicating that the use of three different treadmills did not differentially affect the output from the accelerometers. The treadmill speed was also tested and no differences were identified for speed across trials or treadmills. Because there was limited variability attributed to genders or treadmills, emphasis in the remainder of the analyses was placed on the variability among the different units for each of the four monitors studied.

The descriptive data in Table 2 reveal the variability between the multiple units of a given monitor. The CV values ranged from 7.7 to 20.1% for the four different monitor types (see Table 2). The CV values were similar for the CSA/MTI, Tritrac, and Biotrainer Pro with mean CV across three trials of 8.9, 9.4, and 10.0%, respectively. The mean CV values were considerably higher for the Actical monitor (mean CV across three trials = 20.0%). There were no apparent differences in responses for males and females, but data are reported separately in Table 2 for comparison purposes. The graphs in Figure 1 reveal the individual variability across the different units for each of the monitor types.

Although there was considerable variability across units, the results reveal greater variability when comparisons were made among individuals for a given unit. For these comparisons, the mean counts were computed for each unit in each trial. The SD and CV reveal the variability among individuals when wearing the same monitor and completing the same absolute workload. The mean CV value was computed to reflect the amount of individual variability in these

TABLE 2. Descriptive statistics for the variability in counts across monitors.

Monitor	Variable*	Trial 1 Mean (SD)	Trial 2 Mean (SD)	Trial 3 Mean (SD)
CSA/MTI				
Females (N = 25)	Mean (cnts)	2787 (610)	2789 (526)	2816 (497)
	SD (cnts)	255 (116)	226 (97)	245 (143)
	CV (%)	9.7 (6.1)	8.2 (3.4)	8.8 (4.0)
Males (N = 7)	Mean (cnts)	2923 (374)	3087 (365)	2889 (381)
	SD (cnts)	304 (133)	230 (82)	232 (69)
	CV (%)	10.8 (6.2)	7.7 (3.0)	8.1 (2.7)
Combined (N = 32)	Mean (cnts)	2817 (564)	2854 (506)	2832 (470)
	SD (cnts)	266 (119)	227 (93)	242 (129)
	CV (%)	10.0 (5.9)	8.1 (3.3)	8.6 (3.7)
Biotrainer Pro				
Females (N = 23)	Mean (cnts)	4.75 (0.88)	4.83 (0.66)	4.75 (0.78)
	SD (cnts)	0.44 (0.18)	0.47 (0.13)	0.45 (0.11)
	CV (%)	9.6 (4.3)	9.7 (2.4)	9.6 (2.9)
Males (N = 11)	Mean (cnts)	4.60 (0.72)	4.62 (0.81)	4.83 (0.84)
	SD (cnts)	0.47 (0.16)	0.46 (0.17)	0.54 (0.13)
	CV (%)	10.4 (3.8)	9.8 (2.6)	11.7 (4.4)
Combined (N = 34)	Mean (cnts)	4.70 (0.82)	4.76 (0.71)	4.78 (0.79)
	SD (cnts)	0.45 (0.18)	0.46 (0.14)	0.48 (0.13)
	CV (%)	9.9 (4.1)	9.7 (2.4)	10.3 (3.5)
Actical				
Females (N = 26)	Mean (cnts)	2072 (410)	2240 (834)	2262 (530)
	SD (cnts)	422 (176)	466 (269)	436 (183)
	CV (%)	20.2 (7.3)	20.6 (8.3)	19.2 (5.9)
Males (N = 12)	Mean (cnts)	1745 (203)	1887 (320)	1879 (350)
	SD (cnts)	301 (131)	406 (104)	397 (255)
	CV (%)	17.5 (8.1)	21.6 (4.8)	20.8 (11.5)
Combined (N = 38)	Mean (cnts)	1969 (387)	2129 (727)	2141 (509)
	SD (cnts)	384 (171)	448 (230)	423 (205)
	CV (%)	19.4 (7.6)	20.9 (0.07)	19.7 (7.9)
Tritrac				
Females (N = 22)	Mean (cnts)	1510 (193)	1569 (206)	1614 (197)
	SD (cnts)	143 (55)	155 (62)	150 (53)
	CV (%)	9.4 (3.3)	9.8 (3.6)	9.3 (3.2)
Males (N = 11)	Mean (cnts)	1376 (183)	1394 (184)	1392 (204)
	SD (cnts)	97 (39)	152 (58)	135 (40)
	CV (%)	7.0 (2.4)	10.8 (3.8)	9.9 (3.4)
Combined (N = 33)	Mean (cnts)	1466 (198)	1511 (213)	1540 (223)
	SD (cnts)	128 (54)	154 (59)	145 (49)
	CV (%)	8.6 (3.2)	10.1 (3.6)	9.5 (3.2)

Cnts, raw accelerometer counts.
* The mean, SD, and CV reflect the variability across the different units for each monitor type. The descriptive statistics are reported separately for each of the three trials and the reported values are the mean and SD of these responses across participants.

responses. The mean CV values for the four different monitors were as follows: CSA/MTI (20.1%), Biotrainer Pro (18.1%), Actical (31.1%), and Tritrac (15.9%). There was similar variability across all of the units of a given monitor type, indicating that this variability is fairly typical for each unit. Overall, these results suggest that accelerometer counts for a standardized bout of activity can vary from 16 to 31% for participants performing the same absolute workload on a treadmill.

Generalizability results. The G-study results are presented in Table 3 for the four different monitors. In each case, the largest source of variance in the model is subjects, typically accounting for about 50% of the variance (see Table 3). The highest subject variance was found for the CSA/MTI (63.4%) and the lowest for the Actical (38.5%). The next largest source of variance was for the trial by subject (T × S) interaction term. The values for this term ranged from 14.5% for the CSA/MTI to 21.0% for the Biotrainer with the mean variance being 17.4%. The Monitor term reveals the amount of variance across the multiple units of a given monitor. The smallest amount was found for the CSA/MTI (0.9%) and the largest amount for the Actical (11.7%). The monitor by subject (M × S) term was negligible (~1%) for most monitors except for the Actical (7.4%). This error term reflects differential responses of subjects for the different Actical units and is the most troubling component as it is a more systematic form of error than random error distributed equally across subjects.

The G coefficients in Table 3 provide an overall indication of the reliability of each monitor. For this computation (D-study phase), estimates are based on the use of a single monitor because this is the way that most activity monitors are used (i.e., participants wearing a single monitor over a span of days). The G coefficient for a single trial with a single monitor was highest for the CSA/MTI monitor ($G = 0.64$; SEM = 348). The corresponding values for the other monitors were as follows: Tritrac ($G = 0.573$, SEM = 184), Biotrainer Pro ($G = 0.557$, SEM = 0.664), and Actical ($G = 0.432$, SEM = 557). No G coefficients are reported for the monitors across three different trials or across all 10 monitors because it is unlikely that investigators would ever use this type protocol in subsequent field research (i.e., performing several trials with the same monitor).

Because ICC values are more commonly used in this type of research, we provide estimates using this parameter for comparison purposes. For these computations, we collapsed

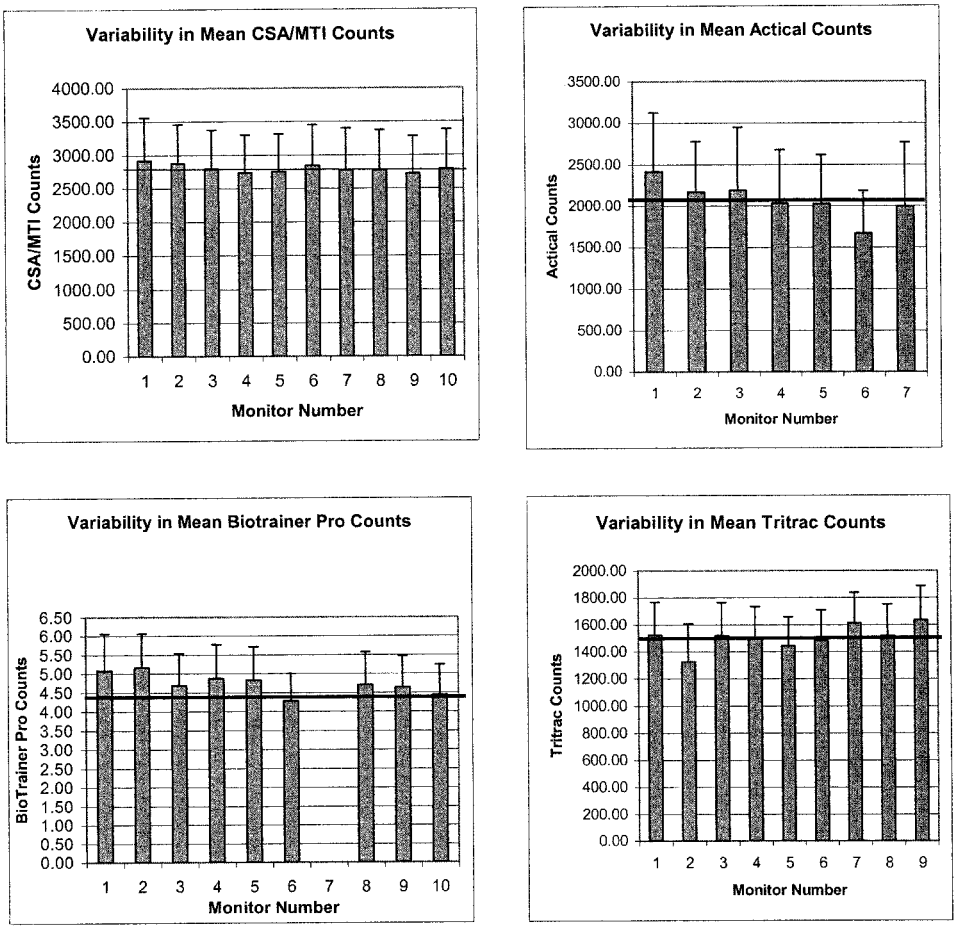

FIGURE 1—Variability in mean activity counts for a standardized bout of activity.

the variability across monitors by using the mean value across the multiple units of a given monitor. The ICC values across the three trials were 0.80 for the CSA/MTI, 0.73 for the Tritrac, 0.68 for the Biotrainer, and 0.62 for the Actical. These indicators of reliability are higher than the G values reported because they do not capture the inherent variability across the different units of each monitor.

Correlational results. Correlations were computed with various anthropometric variables to examine possible influences of body size on accelerometer output. Correlations with height and weight were generally low and nonsignificant but there were some exceptions (see Table 4). There were significant negative correlations with height ($r = -0.54$) and weight ($r = -0.38$) for the combined Tritrac data. The relationships with height were significant for the females ($r = -0.46$) but not for males. Conversely, a significant correlation with height was observed for males wearing the Biotrainer Pro, but this was not true for females. The low magnitude of the correlations and the lack of any systematic pattern for males and females suggest that the observed values may simply be due to chance or to the distribution of the data. In general, there does not appear to be much of a relationship between anthropometric variables and accelerometer counts during treadmill walking with this sample.

DISCUSSION

It is a well-established tenet of science that reliability is a necessary prerequisite for validity, yet little work has been done to examine sources of error in accelerometry-based activity monitors. Most research with accelerometers has emphasized convergent validity against criterion measures of activity and/or the development of calibration studies that establish cutpoints to define different levels of physical activity. It is equally important to establish reliability and identify potential sources of error because these sources impose a natural limit for the validity of these devices. Because free-living activity is inherently variable, the only way to determine reliability is with the use of standardized movement tests. Some studies have tested the reliability of monitors using variance mechanical setups, but this is the first study that has systematically examined sources of variability in activity monitor output under real world conditions (with actual participants performing a standardized bout of physical activity). Because the study evaluated four different monitors using the same protocol, the results provide valuable information about the relative reliability of the different devices.

An additional advantage of this study is that it employs generalizability theory, which allows variance in accelerometer counts to be partitioned and quantified. In activity

TABLE 3. Variance components, variance, and summary reliability statistics for the CSA/MTI, Biotrainer Pro, Actical, and Tritrac R3D monitors.

Source of Variation	CSA/MTI (δ^2)	Biotrainer (δ^2)	Actical (δ^2)	Tritrac (δ^2)
Variance components (δ^2) for each monitor				
Monitors (M)	3079.32	0.08	48738.68	7866.23
Trials (T)	0.00	0.00	4586.37	1054.53
M × T	0.00	0.00	920.36	78.93
Subjects (S)	209894.40	0.42	195152.43	33476.94
M × S	3629.42	0.01	37264.06	58.52
T × S	48094.86	0.18	99478.89	9708.61
M × T × S	66451.57	0.15	120156.80	15184.15
Total	331149.58	0.84	506297.60	67427.92
Source of Variation	**CSA/MTI % Variance**	**Biotrainer % Variance**	**Actical % Variance**	**Tritrac % Variance**
Variance associated with each component (%) for each monitor				
Monitors (M)	0.9	9.4	9.6	11.7
Trials (T)	0.0	0.0	0.9	1.6
M × T	0.0	0.2	0.2	0.1
Subjects (S)	63.4	50.4	38.5	49.6
M × S	1.1	1.2	7.4	00.1
T × S	14.5	21.0	19.6	14.4
M × T × S	20.1	17.8	23.7	22.5
Total	100.0	100.0	100.0	100.0
Summary statistics for each monitor				
G	0.640	0.557	0.432	0.573
SEM	348.217	0.644	557.804	184.258
ICC	0.80	0.68	0.62	0.73

G, generalizability coefficient; ICC, intraclass correlation coefficient. Note: G coefficients are reported for a single instrument (unit) on one trial since this is the way that monitors are used in practice.

monitoring research the goal is to understand differences in activity levels among participants. If the monitors were completely free from error, the subject term would account for 100% of the variance and there would be no variability due to the other components (e.g., between trials or between monitors). The subject term accounted for the largest percent of the variance for all monitors with the largest percentage amount for the CSA/MTI monitor. Other variance components accounted for the remaining variance and these components contribute unwanted error in activity monitoring research.

The variability due to differences across monitoring units is captured with the monitor (M) term. The CSA/MTI had the smallest M component (0.9%) followed by the Biotrainer Pro (9.4%), Actical (9.6%), and the Tritrac (11.6%). This error term may be influenced by quality control factors in manufacturing or by the unique acceleration response of the component accelerometer within the device. Brage et al. (4) reported significant differences in the filtering within individual accelerometer devices, so this type of error may account for some of the variability. Another possible source of error for the M component would be differences in monitor calibration. In the present study, the CSA/MTI and Actical units were brand new, whereas the Biotrainer Pro and Tritrac units had been previously used for some field based monitoring in previous research. The Biotrainer Pro monitors were sent back to the manufacturer (IM Systems) to check calibration following the testing, and all of the monitors were found to be within the manufacturer specifications. Therefore, the previous use of the Biotrainers did not appear to alter their calibration. The Tritrac R3D is no longer manufactured and the current owner of the technology (Stayhealthy.com) does not provide support for these devices, so it was not possible to obtain a recalibration check with this device. The website describing the new RT3 monitor pointed out weaknesses in quality control procedures (lack of preproduction testing of the accelerometers and soldering of individual accelerometers within the unit) as limitations of the original Tritrac R3D models. These factors may have contributed to the larger amount of variance accounted for by the M term for the Tritrac monitor. The fact that the new Actical units had similar M values suggests that previous use may not be a major factor.

The trial (T) term reveals the variability across trials for each of the monitors. This was essentially negligible for all of the monitors as it accounted for less than 2% of the total variation. The low variance accounted for this term suggests that the monitors provide similar information over time for a standardized bout of activity. While the trial term is minimal, the trial by subject (T × S) interaction term accounted for a sizable amount of variance in the data (range: 14–21%). This term reflects the variability associated with a participant's results across trials, and the sizable interaction component suggests that there are differential effects across the subjects. If monitors are positioned or angled in slightly different ways across individuals or between trials for a given individual, measurement error would be added to the accelerometer signal or output. Error in the overall trial level effect would tend to average out, but the T × S term would capture the individual variability in this response.

Specific efforts were made in the study to minimize variability due to positioning. The monitors were positioned to align with marked spots on the belt and research assistants helped the participants maintain a similar orientation within and between trials. The relatively high amount of variance accounted for by the T × S and the M × T × S components suggests that significant variability occurs across trials and that this difference varies by participant (possibly due to clothing, posture, or other anthropometric variations). These differences are likely to be greater when monitors are worn under free-living conditions because it is even harder to standardize positions on the hip and maintain a proper orientation of the monitor during field-based monitoring over several days.

Output from uniaxial devices such as the CSA/MTI, Biotrainer would be expected to be more susceptible to differences in positioning as the magnitude of the acceler-

TABLE 4. Correlations between activity counts and anthropometric variables.

Monitor	Sample 1 CSA/MTI	Sample 2 Biotrainer Pro	Sample 3 Actical	Sample 4 Tritrac R3D
Females	$N = 25$	$N = 23$	$N = 26$	$N = 22$
Height (m)	0.24	−0.10	0.08	−0.46*
Weight (kg)	0.16	0.35	−0.15	−0.24
BMI	−0.01	0.42*	−0.19	−0.09
Males	$N = 7$	$N = 11$	$N = 12$	$N = 11$
Height (m)	−0.28	−0.64*	0.01	−0.15
Weight (kg)	−0.37	−0.14	0.02	−0.03
BMI	−0.24	0.06	0.02	0.05
Combined	$N = 32$	$N = 34$	$N = 38$	$N = 33$
Height (m)	0.23	−0.22	−0.18	−0.54*
Weight (kg)	0.16	0.06	−0.29	−0.38*
BMI	0.08	0.22	0−0.28	−0.17

* $P < 0.05$.

ation signal depends on the relative positioning with respect to gravity. In contrast, the three-dimensional output from the Tritrac (vector magnitude) should be relatively immune from this type of effect; the direction of the vector in three-dimensional space could vary but the magnitude should theoretically be the same regardless of position. This hypothesis was supported to some extent as the Tritrac exhibited the lowest T × S component of the four monitors. The Biotrainer Pro had a larger T × S term than the CSA/MTI and this may reflect the less stable method of attachment used for the Biotrainer (belt clip) compared with the CSA/MTI (attached to a belt). The omni-directional sensor in the Actical would be expected to be similar to the Tritrac in performance but this device had the largest T × S term. Information on the technical specifications and functionality of the "omni-directional" sensor may help to explain this type of effect.

A more troubling error component is the monitor × subject term (M × S). This term captures differential variability in responses of participants across the individual monitoring units. This term was negligible for the CSA/MTI, Biotrainer, and the Tritrac but was larger (7.4%) for the Actical. It is not clear what might have caused this response for the Actical, but it influenced the overall reliability of the device as the Actical had the smallest overall G value of the four monitors. The Tritrac had the smallest variance due to this term, suggesting again that the three-dimensional output (vector magnitude) may help to reduce variability of this type.

Overall, the study provides valuable quantitative data on the sources of variability in accelerometry-based activity monitors. The computed G values were highest for the CSA/MTI monitor and lowest for the Actical monitor. The CSA/MTI monitor has become the most widely used and accepted monitor among researchers in the field, and the results of this study provide support for its continued use. The main difference between the CSA/MTI and the other monitors was in the monitor (M) component. As mentioned, the CSA/MTI units were brand new while the Biotrainer Pro and Tritrac units were previously used, so the findings here may have been influenced to some extent by the past use (even though they were still calibrated). The poor reliability results with the Actical monitor were particularly concerning as these units were brand new. The monitor is very small and of apparent high quality and sophistication, but the results clearly revealed poorer reliability for this monitor. Additional testing is needed with this device before definitive conclusions can be made about its accuracy and reliability.

In general, direct comparisons between monitors must be made with caution. The data collection for the present study required repeated laboratory visits for each participant and were compiled over multiple semesters, so it wasn't possible to use a single sample of participants for all monitors. Unique characteristics of the samples could influence the results, but there were no differences in anthropometric values or demographic measures between the samples, and they were recruited from a similar sample population. The emphasis in the present study was on the variability within individuals and between trials, so the nature of the sample would only influence comparisons between the different monitors. The G values for each device are computed separately for each monitor but comparisons between G values should still be done with caution.

There are some additional limitations in our design that should also be considered when interpreting the results. First of all, the nonrepresentative nature of the sample could have influenced the results to some degree. The participants were exercise science majors and were leaner and generally more fit than the average population, so our results may not generalize to all populations. The small sample of males may also make the data less representative, but gender did not seem to be an important factor in the analyses. The present study evaluated the reliability of these monitors for one standardized bout of activity (treadmill walking), so the results may not generalize to other settings or activities. The reliability under free-living conditions would be expected to be less than under these standardized conditions (for all monitors), but this would be difficult to test because it would not be possible to standardize the activity as in the present study.

The documentation of considerable variability in accelerometer units has important implications for the design of studies using activity monitors to assess free-living physical activity. There has recently been greater interest in using activity monitors as primary outcomes in intervention studies. Because these designs involve the use of multiple monitors on large samples of people at multiple time points, it is important to control for as much error as possible. Estimates of these variance components for power analyses and greater attention to positional influences (3) may help to improve the accuracy of these devices for this type of research. Greater attention is also needed on issues regarding the calibration of these devices (4). A calibration device can be purchased for use with the CSA/MTI, but comparable devices are not available for other monitors. The website on the Actical monitor specifically indicates that recalibration is not necessary but assuring quality control in all facets of data collection would likely enhance the quality of the data that is obtained with these monitors. Maintaining calibration would be particularly important in studies in which repeated measurements are taken over time or in tracking studies in which data are collected in multiple years. Future work should systematically study this issue. The use of similar designs and analytical techniques (generalizability theory) offer considerable promise for advancing the knowledge base on factors influencing output from accelerometry-based activity monitors.

The authors wish to acknowledge the contributions of Research Assistants (Sara Neuhaus and Amber Long) for their contributions in collecting this data.

REFERENCES

1. BASSETT, D. R., Jr., B. E. AINSWORTH, A. M. SWARTZ, S. J. STRATH, K. O. O'BREIN, and G. A. KING. Validity of four motion sensors in measuring moderate intensity physical activity. *Med. Sci. Sports Exerc.* 32:S471–S480, 2000.
2. BOUTEN, C. V. C., K. T. M. KOEKKOEK, M. VERDUIN, R. KODDE, and J. D. JANSSEN. A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. *IEEE Trans. Biomed. Eng.* 44:136–147, 1997.
3. BOUTEN, C. V. C., SAUREN AAHJ, VERDUIN M., JANSSEN J. D. Effects of placement and orientation of body-fixed accelerometers on the assessment of energy expenditure during walking. *Medical and Biological Engineering and Computing.* 35:50–6, 1997.
4. BRAGE, S., N. BRAGE, N. WEDDERKOPP, and K. ROBERG. Reliability and validity of the computer science and applications accelerometer in a mechanical setting. *Measurement in Phys. Education and Exerc. Science.* 7:101–19, 2003.
5. COLEMAN, K. J., B. E. SAELENS, M. D. WIEDRICH-SMITH, J. D. FINN, and L. H. EPSTEIN. Relationships between Tritrac-R3D vectors, heart rate, and self-report in obese children. *Med. Sci. Sports Exerc.* 29:1535–1542, 1997.
6. ESTON, R. G., A. V. ROWLANDS, and D. K. INGLEDEW. Validity of heart rate, pedometry, and accelerometry for predicting the energy cost of children's activities. *J. Appl. Physiol.* 84:362–371, 1998.
7. FINN, K. J., and B. SPECKER. Comparison of Actiwatch activity monitor and Children's Activity Rating Scale in children. *Med. Sci. Sports Exerc.* 32:1794–1797, 2000.
8. FREEDSON, P. S., E. MELANSON, and J. SIRARD. Calibration of the Computer Science and Applications, Inc. Accelerometer. *Med. Sci. Sports Exerc.* 30:777–781, 1998.
9. GOODWIN, L. D. Interrater agreement and reliability. *Meas. Phys. Educ. Exerc. Sci.* 5:13–34, 2001.
10. HENDLEMAN, D., K. MILLER, C. BAGGET, E. DEBOLD, and P. S. FREEDSON. Validity of accelerometry for the assessment of moderate intensity physical activity in the field. *Med. Sci. Sports Exerc.* 32:S442–S449, 2000.
11. JAKICIC, J. M., C. WINTERS, K. LAGALLY, J. HO, R. J. ROBERTSON, and R. R. WING. The accuracy of the Tritrac-R3D accelerometer to estimate energy expenditure. *Med. Sci. Sports Exerc.* 31:747–754, 1999.
12. KOCHERSBERGER, G., E. MCCONNELL, M. N. KUCHIBHATLA, and C. PIEPER. The reliability, validity, and stability of a measure of physical activity in the elderly. *Arch. Phys. Med. Rehabil.* 77:793–795, 1996.
13. LEENDERS, N. Y. J. M., W. M. SHERMAN, and H. N. NAGARAJA. Comparison of four methods of estimating physical activity in adult women. *Med. Sci. Sports Exerc.* 32:1320–1326, 2000.
14. METCALF, B. S., J. S. CURNOW, C. EVANS, and L. D. W. T. J. VOSS. Technical reliability of the CSA activity monitor: the Early Bird Study. *Med. Sci. Sports Exerc.* 34:1533–1537, 2002.
15. MORROW, J. R. J. Generalizability theory. In: *Measurement Concepts in Physical Education and Exercise Science,* M. J. Safrit and T. Wood (Eds.). Champaign, IL: Human Kinetics, 1989, pp. 73–96.
16. NICHOLS, J. F., C. G. MORGAN, J. A. SARKIN, J. F. SALLIS, and K. J. CALFAS. Validity, reliability, and calibration of the Tritrac accelerometer as a measure of physical activity. *Med. Sci. Sports Exerc.* 31:908–912, 1999.
17. SIRARD, J. R., E. L. MELANSON, and P. S. FREEDSON. Field evaluation of the Computer Science and Applications, Inc. physical activity monitor. *Med. Sci. Sports Exerc.* 32:695–700, 2000.
18. TROST, S. G., D. S. WARD, and J. R. BURKE. Validity of the Computer Science and Application (CSA) Activity Monitor in children. *Med. Sci. Sports Exerc.* 30:629–633, 1998.
19. WELK, G. J. Use of accelerometry-based activity monitors to assess physical activity. In: *Physical Activity Assessments for Health Related Research,* G. J. Welk (Ed.). Champaign, IL: Human Kinetics, 2002, pp. 125–141.
20. WELK, G. J., J. ALMEIDA, and G. MORSS. Laboratory calibration and validation of the Biotrainer and Actitrac activity monitors. *Med. Sci. Sports Exerc.* 35:1057–1064, 2003.
21. WELK, G. J., S. N. BLAIR, K. WOOD, S. JONES, and K. W. THOMPSON. A comparative evaluation of three accelerometry-based physical activity monitors. *Med. Sci. Sports Exerc.* 32: S489–S497, 2000.
22. WELK, G. J., and C. B. CORBIN. The validity of the Tritrac-R3D activity monitor for the assessment of physical activity in children. *Res. Q. Exerc. Sport* 66:202–209, 1995.