# Protocols for Evaluating Equivalency of Accelerometry-Based Activity Monitors

GREGORY J. WELK[1], JAMES MCCLAIN[2], and BARBARA E. AINSWORTH[3,4]

[1]*Department of Kinesiology, College of Human Sciences, Iowa State University, Ames, IA;* [2]*Risk Factor Monitoring and Methods Branch, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD;* [3]*Healthy Lifestyles Research Center, College of Nursing and Health Innovation, Arizona State University, Phoenix, AZ; and* [4]*Program in Exercise and Wellness, College of Nursing and Health Innovation, Arizona State University, Phoenix, AZ*

**ABSTRACT**

WELK, G. J., J. MCCLAIN, and B. E. AINSWORTH. Protocols for Evaluating Equivalency of Accelerometry-Based Activity Monitors. *Med. Sci. Sports Exerc.*, Vol. 44, No. 1S, pp. S39–S49, 2012. A wide array of accelerometer-based activity monitors has been developed to facilitate objective monitoring of physical activity behaviors, but it has proven difficult to equate outputs from different monitors. On the surface, commercially available monitors seem to be performing the same basic task—monitoring total body acceleration. However, differences in sensor properties and internal data processing have made it difficult to directly compare output from different monitors. In recent years, many new competing technologies have been released into the market, compounding the challenge of evaluating monitor equivalency and the relative strengths and limitations of different monitors. To advance physical activity assessment and improve our ability to compare results across studies using different monitors, it is important to conduct functional equivalency studies in a standardized and systematic way. This article summarizes issues associated with monitor equivalency and proposes methods for standardization and quality control in future research. **Key Words:** ACCELEROMETERS, ASSESSMENT, PHYSICAL ACTIVITY, ACTIVITY MONITORS

The use of accelerometry-based activity monitors for evaluating physical activity patterns has become an accepted practice in physical activity research. Although methodological practices and quality control have improved in recent years, many obstacles remain to be overcome (28). The inherent challenges associated with collecting, processing, and interpreting data from activity monitors have been well chronicled in the literature and in this supplement. Historically, less attention has been given to the fundamental differences between monitors and between units of a given monitor type.

Researchers have routinely used the term "accelerometers" to generically refer to accelerometry-based activity monitors. There are a wide range of commercially available monitors produced by various manufacturers. Furthermore, the progression of technology and manufacturers' responsiveness to interest in specific measurement functions has resulted in availability of multiple monitor models from the same company. It is clear that the various commercially available monitors are inherently different (8). If all monitors were truly accelerometers, they could be used relatively interchangeably to provide data on body acceleration. Although most activity monitors are based on acceleration data from internal accelerometers, the internal processing leads to different outputs that cannot be directly compared. Researchers have presumed that results will be equivalent between multiple units of a given activity monitor, but this does not always hold true. Because reliability is a prerequisite of validity, it is important to evaluate monitors on the basis of reliability as well as validity. As stated by Esliger et al. (10), "the quality of information from accelerometers is only as good as the devices themselves."

This article summarizes issues associated with monitor equivalence and provides recommendations for improving standardization and quality control in the future. The first section summarizes past monitor equivalency research and provides a foundation for the review. The second section highlights emerging measurement challenges that complicate current monitor equivalency studies. Emphasis is placed on the inherent challenges of comparing different monitoring technologies and approaches. The article concludes with general guidelines and best practices for monitor equivalency research. This section is intended to facilitate future research aimed at determining the relative utility of different

Address for correspondence: Gregory J. Welk, Ph.D., 257 Forker Building, Ames, IA 50011; E-mail: gwelk@iastate.edu.

0195-9131/12/441S-0S39/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/MSS.0b013e3182399d8f

accelerometry-based activity monitors for assessing free-living physical activity and energy expenditure.

## CURRENT STATE OF RESEARCH

**Can output from two monitors be directly compared?** The output of accelerometry-based devices has been generically termed "counts," but counts cannot be directly compared across monitors because of differences in how the raw data are collected, processed, filtered, and scaled (8). It would seem possible, in theory, to achieve some degree of standardization in output if companies converted the raw signals into real acceleration units (scaled in meters per second squared or Newtonian constant of gravitation). Unfortunately, this is more complicated than it would seem. Research by Brage et al. (6), for example, demonstrated that ActiGraph (Pensacola, FL) output is proportional to movement acceleration only if frequency is held constant. More recently, Esliger et al. (10) reported varying responsiveness to changes in acceleration and frequency in different monitors. Variability in Actical (Mini Mitter Co., Inc., Bend, OR) output was found to be negatively related to acceleration (i.e., increased variability or error at lower accelerations). In contrast, variability in the ActiGraph was found to be negatively related to the frequency of the acceleration signal. Although the two devices had similar performance characteristics overall, the discrepant responsiveness to acceleration and frequency between monitors make it difficult to determine a "superior" model. These results demonstrate that output is dependent on both the frequency and acceleration detected by the monitor. The variations in filtering and scaling cause monitors to function differently when exposed to a common stimulus. The effect of differences in filtering and scaling of acceleration signals is displayed in Figure 1 for the ActiGraph GT1M, Actical, and RT3 (Stayhealthy, Inc., Monrovia, CA) (data provided by Kong Chen and Megan Rothney, personal communication). The disparate results clearly demonstrate why monitor output cannot be directly compared. If similar filtering and scaling methods were used by all companies, standardization might be possible. It is not practical or realistic to expect companies to agree on a common method for processing at this point in the evolution of accelerometry-based monitors. Several companies have recently developed commercial monitors that are capable of logging raw, unfiltered accelerations in three axes for extended periods including ActiGraph LLC; Activinsights, Ltd.; and PAL Technologies, Ltd. (Glasgow, Scotland). However, even raw data monitors will require rigorous equivalency testing to determine whether and under what conditions device outputs can be directly compared.

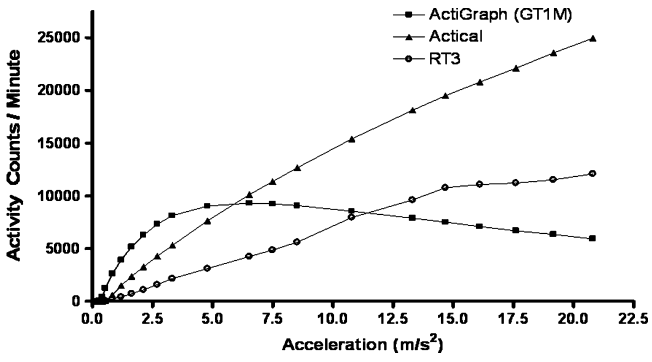


**FIGURE 1—Relationship between accelerations and activity counts-per-minute outputs from the ActiGraph GT1M, Actical, and RT3 in a mechanical setup. Observed difference output response reflects variation in filtering and scaling parameters used by manufacturers.**

**Do two units of the same monitor model produce the same output?** Another key question in accelerometry research is whether different units of the same monitor model provide equivalent information. Researchers have generally been more concerned about the validity of the monitor they choose, but reliability has important implications on the quality of data that are obtained in any study of free-living participants. Reliability has become increasingly important as monitors begin to be used for public health surveillance and in large-scale clinical trials that collect data across multiple time points.

Evaluating the reliability of accelerometer-based activity monitors necessitates exposing multiple units of the same monitor model to the same conditions. Reliability studies have been conducted under both laboratory and free-living conditions, and each approach has advantages and disadvantages. Laboratory studies generally use some type of mechanical device (e.g., turntables, wheels, vibration tables, or shaker tables) to impart a standardized movement or acceleration to multiple units of a monitor model (6,10,14, 21,24,27). These studies have yielded new insights for individual monitors, but few studies to date have compared unit reliability of competing monitors (29). Results of laboratory studies may not generalize to actual reliability when used in studies of free-living individuals. Studies conducted under real-world conditions typically compare output during standardized activities (e.g., walking on a treadmill) or assess the reliability of two monitors positioned on the right and left sides of the body (19,25,26,30,33). These free-living designs simulate natural conditions but are limited by the inability to fully standardize the conditions and positions of the accelerometers. Without this control, it is difficult to determine sources of variability in accelerometer output. Both laboratory and free-living reliability study designs clearly have strengths and limitations. The distinction between study designs can be likened to the relative emphasis placed on internal or external validity criteria in other research designs.

A comprehensive summary of the accelerometer reliability literature is beyond the scope of this review, but some generalizations are provided. Early studies revealed poor reliability for many monitor models, but results from more recent studies have been more favorable. For example, results with newer technology have been far better than earlier devices (e.g., the ActiGraph GT1X model vs ActiGraph 7164 model), and this is likely attributable to improved quality control and

the use of solid-state electronics (27). Research also shows clearly that variability is greater between units than within units of the same monitor model (6,10). This makes sense because the variability between units includes the error that is caused by the inherent differences in positioning on the body or unit functioning that influences within-unit (day-to-day) comparisons. The inherent error between and within units clearly affects the ability to accurately estimate levels of physical activity in the population. However, it is likely that this error is masked by larger sources of variance due to calibration issues and inability to equally assess all types of activity in a consistent way. Although additional research is clearly needed, the reliability of most accelerometry-based monitors is likely acceptable for most research applications.

**Do different monitors provide equivalent information?** As described above, accelerometer output is determined by many factors (type of sensor, filtering procedures, and processing and scaling). Although the magnitude of counts varies in competing monitors, it is possible to make direct comparisons between monitors on the basis of the relative ability to estimate time spent being active or amount of energy expenditure. However, these comparisons are complicated because the outcome measures are influenced by both the raw signals and the inherent differences in prediction equations or algorithms.

Many studies have been conducted to compare monitor function and performance. An early study by Welk et al. (31) directly compared three monitors (ActiGraph, BioTrainer (IM Systems, Baltimore, MD), and Tritrac (Stayhealthy, Inc.)) under both laboratory and free-living conditions to determine the comparability of their output. The monitors provided reasonably accurate estimates of energy expenditure for locomotor activity on a treadmill, but all monitors were shown to underestimate the energy cost of the free-living activities determined by indirect calorimetry. That study pointed out some limitations in assessing diverse types of activities and the tendency to underestimate energy expenditure for lifestyle tasks that involve a lot of upper body movement. A key observation from that article is that correlations between monitors were high under both laboratory ($r$ = 0.86) and field ($r$ = 0.70) conditions, indicating that the monitors provide similar information under both conditions.

Table 1 provides a summary of some recent studies that have directly compared different monitors under laboratory or free-living conditions. The majority of these studies have compared two or three different monitors during a period of 1 to 7 d in free-living settings. The majority of studies identified included the ActiGraph model 7164 or GT1M. Comparisons were made between counts (counts per minute), energy expenditure, or time spent in light-, moderate-, and vigorous-intensity activities. The analytic methods used in recent comparison studies have predominantly used tests of overall associations (e.g., correlations) and group means (e.g., $t$-tests or ANOVA), with more limited use of methods for testing of potential bias (e.g., Bland–Altman plots). The diverse nature of the studies and the variability in results make it difficult to draw conclusions about agreement among monitors. Studies also are limited because of small (10 to

TABLE 1. Overview of instruments, methods, and participant samples used in recent monitor comparison studies.

| Study | Sample | Instruments | Setting (Duration) | Outcome Measures | Analytic Methods |
|---|---|---|---|---|---|
| Adults: | | | | | |
| Macfarlane et al. (17) | $n$ = 49, M and F<br>Age = 28.6 ± 9.0 yr | AG-7164<br>RT3 | Free-living (7 d) | Light<br>Moderate<br>Vigorous | Correlations<br>Friedman nonparametric tests |
| Welk et al. (32) | $n$ = 30, M and F<br>Age = 24.9 ± 6.1 yr | AG-7164<br>SenseWear Pro2<br>IDEEA | Free-living (1 d) | EE<br>MVPA | Correlations<br>ANOVA<br>Bland–Altman |
| McClain et al. (18) | $n$ = 10, M and F<br>Age = 29.0 ± 2.3 yr | AG-7164<br>Lifecorder EX | Free-living (1 d) | Light<br>MVPA | Correlations<br>ANOVA |
| Paul et al. (22) | $n$ = 56, M and F<br>Age = 30–60 yr | AG-7164<br>Actical | Free-living (15 d) | Counts per minute | Correlations<br>CV of differences<br>Bland–Altman<br>Regression |
| Abel et al. (1) | $n$ = 20<br>Age = 29.4 ± 7.1 yr | AG-GT1M<br>Lifecorder EX | Free-living (1 d) | Moderate<br>Vigorous<br>Steps | Correlations<br>ANOVA |
| Herrmann et al. (12) | $n$ = 16<br>Age = 40.2 ± 12.6 yr | AG-GT1M<br>Active Key | Laboratory/free-living (7 d) | Light<br>Moderate<br>Vigorous | Correlations |
| McClain et al. (18a) | $n$ = 26, M and F<br>Age = 27.3 ± 7.1 yr | AG-GT1M<br>NL-1000<br>Omron HJ-151 | Free-living (1 d) | MVPA | Correlations<br>ANOVA |
| Children: | | | | | |
| Corder et al. (9) | $n$ = 30, M and F<br>Age = 15.8 ± 0.6 yr | AG-7164<br>AG-GT1M | Free-living (7 d) | Sedentary<br>Light<br>Moderate<br>Vigorous<br>Counts per minute | Correlations<br>ANOVA<br>Bland–Altman |
| McClain et al. (20) | $n$ = 31, M and F<br>Age = 10.2 ± 0.4 yr | AG-7164<br>Lifecorder EX | Free-living (1 d) | Light<br>MVPA | Correlations<br>ANOVA |

F, female; M, male; MVPA, moderate to vigorous physical activity; EE, energy expenditure.

50 persons) and homogeneous (similar age, adiposity, and activity levels) convenience samples. Although these studies have contributed new insights, it still is not possible to determine the relative utility of monitors or the relative degree of correspondence for evaluating physical activity or energy expenditure. Recommended guidelines for comparison studies are provided at the end of the article to guide future research.

## EMERGING MEASUREMENT CHALLENGES

In the early years of accelerometry research, only a few monitors were available, and the overall differences in technology were relatively minor. The landscape has changed drastically in the past 5 yr, and many competing technologies are now available for objective physical activity monitoring. The expansion and diversity of technology have made it difficult to determine the relative value or utility of these various tools. Standard accelerometry-based activity monitors have become more sophisticated, and these advances have allowed companies to increase accuracy and functionality while maintaining or reducing cost. However, continuing changes in hardware and firmware (programs that internally control electronic devices) have made it difficult to determine how to compare output even across generations of monitor models from the same manufacturer. The commonly used ActiGraph monitor, for example, has undergone numerous transformations over the years. Considerable efforts have been made to ensure consistency in output across generations, yet research has demonstrated some differences in the sensitivity and properties of the sensors when used for assessing physical activity (9,27).

Advanced technologies and modeling techniques also have led to the development of new pattern recognition devices that provide alternative ways of measuring and evaluating physical activity. These devices use different inputs and may yield outputs different from those of standard accelerometry-based devices. Because of these contrasts, it has proven difficult to determine the relative advantages of the different approaches. The SenseWear Armband (BodyMedia, Inc., Pittsburgh, PA), for example, uses accelerometers in combination with various heat-related sensors to estimate energy expenditure (13). The device incorporates accelerometer data, but the composite indicator of energy expenditure is based on more than raw accelerometer counts. The activPAL (PAL Technologies, Ltd.) is another new device that uses a pattern recognition approach for estimating activity levels (11). The activPAL monitor uses a uniaxial accelerometer with sensitivity to both static and dynamic acceleration to detect postures (sitting/lying, standing) and periods of walking. The Intelligent Device for Energy Expenditure and Activity (MiniSun LLC, Fresno, CA) monitor operates similarly and integrates data from five sensors to detect postures and then applies standard prediction equations to estimate energy cost (34). These devices are highlighted as examples of the diverse ways in which physical activity can be assessed. Because contemporary devices use principles and assumptions different from those of traditional accelerometry-based monitors, it is difficult to make direct comparisons.

The availability of low-cost accelerometer technology also has spurred the development of many small high-quality monitors for consumer use. Other lower priced technologies, such as pedometers, have added features that, in some instances, allow them to provide a functionality that is similar to or better than that of more advanced accelerometry-based activity monitors (2). Unfortunately, these new monitoring devices have been released into the market with impressive claims but little supporting evidence regarding performance. This convergence of technology and function presents an emerging measurement challenge because it further complicates the inherent challenge of comparing and standardizing data from accelerometry-based activity monitors.

Table 2 provides a summary of the different types of monitoring technologies that are currently available. Monitors vary considerably in cost and sophistication. A major challenge for the field is to establish procedures so the relative strengths or limitations of new monitors can be directly and effectively compared. It is likely that the commercial nature of these instruments will drive an even greater range of features and options in the future, increasing both the complexity and challenge of monitor equivalency research. Best practices for evaluating functional equivalence are described in the following section.

## BEST PRACTICES FOR EVALUATING FUNCTIONAL EQUIVALENCE

The previous sections summarized some measurement issues related to monitor equivalency. The emphasis was on evaluating and establishing agreement between monitor types as well as between units of a given monitor. Many studies have been done to examine these issues, but additional work is clearly needed. Research with accelerometers would improve and progress more systematically if studies followed some common guidelines or best practices. Research also would progress more systematically if investigators were expected to demonstrate possible advantages of new monitors relative to available technology. From this perspective, demonstrating that a new monitor can measure activity or energy expenditure is not sufficient unless the researchers can also demonstrate how it compares with or improves upon other currently available techniques (i.e., does the new monitor provide advantages in terms of validity, reliability, utility, feasibility, or cost?). To address this need, researchers are strongly encouraged to use multiple monitors in validation or cross-validation studies so that direct comparisons can be made between two or more competing technologies. In medical research, investigators routinely compare treatments with the "standard-of-care" approach. Accelerometer research would progress more

TABLE 2. Summary of sensor type, data outputs, technical specifications, and cost for a selection of commercially available monitors.

| Instrument (Selected Distributor Web Sites) | Sensor Mechanism | Raw Data Outputs | Minimum Epoch Length/Output Data Aggregation | Research or User Data Interface | PS | BL | MEM | Cost (US$) December 1, 2009 |
|---|---|---|---|---|---|---|---|---|
| Traditional accelerometry-based activity monitors: | | | | | | | | |
| ActiGraph GT3X (*www.the*actigraph.com) | Triaxial accelerometer | 1) Activity counts for *x*, *y*, and *z* axes<br>2) Steps | 1 s or raw data at 32 Hz/by epoch | PC-based software | RLIB | 20 d | 4 GB | 335 |
| Actical (actical.respironics.com) | Omnidirectional accelerometer | 1) Activity counts: single channel<br>2) Steps | 15 s/by epoch | PC-based software | CCLB | 180 d | 64 KB | 450 |
| BioTrainer Pro (*www.imsystems.net*) | Biaxial accelerometer | 1) Activity counts: single channel<br>2) Estimated PAEE | Variable epochs (15 s to 1 min) and sensitivities | PC-based software | AAA battery | 6 months | 112 d | 200 |
| RT3 (*www.stayhealthy.com*) | Triaxial accelerometer | 1) Activity counts for *x*, *y*, and *z* axes<br>2) Steps | 1 s/by epoch | PC-based software | AAA battery | 60 d | 3 h to 7 d | 300 |
| Kenz Lifecorder EX (*www.new-lifestyles.com*) | Single-axis accelerometer | 1) Steps<br>2) Frequency counts for each of 11 categories | 4 s/by day | PC-based software | CCLB | 5 months | 200 d | 245 |
| Pattern recognition physical activity monitors: | | | | | | | | |
| SenseWear Armband (*www.bodymedia.com*) | Triaxial accelerometer, skin temperature, heat flux, GSR sensor | 1) Acceleration, heat, and GSR data for 13 unique sensor channels<br>2) Steps | 1 s/by epoch | PC-based software | RLIB | 7 d | 2 MB (14 d) | 950 |
| IDEEA (*www.minisun.com*) | Multiple accelerometer sensor array | 1) Raw acceleration data at 32 Hz for eight sensor channels<br>2) Posture and gait | 32 Hz/by epoch | PC-based software | AA battery | 60 h | 200 MB | 2995 |
| Actiheart 4 (*www.camntech.com*) | Omnidirectional accelerometer with HR via ECG electrode | 1) Activity counts: single channel<br>2) HR | 15 s/by epoch | PC-based software | RLIB | 21 d | 4–21 d | 1250 |
| ActivPAL (*www.paltech.plus.com*) | Single-axis accelerometer | 1) Time sitting/lying, standing, and walking<br>2) Steps<br>3) Posture and gait | 15 s/by epoch | PC-based software | RLIB | 8 d | 8 d | 950 |
| ActiTrainer (research model) (*www.the*actigraph.com) | Triaxial accelerometer with Polar chest strap | 1) Activity counts for *x*, *y*, and *z* axes<br>2) Steps<br>3) HR | 10 s/by epoch | PC-based software | RLIB | 14 d | 4 MB | 465 |
| StepWatch activity monitor (orthocareinnovations.com) | Ankle-worn custom accelerometer-based sensor | 1) Strides | 3 s/by epoch | PC-based software | RLIB | 7 yr | 30 d at 1-min epoch | 525 |
| DynaPort MiniMod (*www.mcroberts.nl*) | Triaxial accelerometer with optional gyroscopic sensors | 1) Raw acceleration data for *x*, *y*, and *z* axes<br>2) Posture, posture transitions, and gait | 1 s/by epoch | PC-based software | RLIB or AAA battery | 1–3 d | Scalable with SD memory card | 1500 |
| Consumer-based physical activity monitors: | | | | | | | | |
| DirectLife (*www.directlife.philips.com*) | Triaxial accelerometer | No raw data accessible: summary measures include estimated TDEE and time in walking and running activities | 60 s/data by day and graphics by day and hour | Web-based data server with personal summaries | RLIB | 21 d | 21 d | 99 + 12.50 (monthly Web site fee) |

*(continued on next page)*

TABLE 2. *(Continued)*

| Instrument (Selected Distributor Web Sites) | Sensor Mechanism | Raw Data Outputs | Minimum Epoch Length/Output Data Aggregation | Research or User Data Interface | PS | BL | MEM | Cost (US$) December 1, 2009 |
|---|---|---|---|---|---|---|---|---|
| Bodymedia Fit (*www.bodymedia.com*) | Triaxial accelerometer, skin temperature, Heat Flux, GSR sensor | No raw data accessible: summary measures include steps, estimated TDEE, time in specific activity intensities, and sleep measures | 60 s epochs data by day and graphics by day, hour, or minute | Web-based data server with personal summaries | RLIB | 7 d | 1 MB (14 d) | 200 + monthly Web site fee |
| MiBand (www.cambridgeconsultants.com) | Triaxial accelerometer | No raw data accessible: summary measures include steps, estimated TDEE, and time in specific activity intensities | 60 s epochs data by day and graphics by day, hour, or minute | Web-based data server with personal summaries | RLIB | 3 wk | 14 d | 69.12 + 14.90 (monthly Web site fee) |
| Fitbit (*www.fitbit.com*) | Triaxial accelerometer | No raw data accessible: summary measures include steps, estimated TDEE, time in specific activity intensities, and sleep measures | 60 s epochs data by day and graphics by day, hour, or minute | Web-based data server with personal summaries | RLIB | 10 d | 7–30 d | 99 (Web site access included) |
| KAM monitor (also marketed under product name PAM) (*www.mykamunity.com*) | Uniaxial accelerometer | Output in a processed unit that is scaled in terms of MET (i.e., independent of body weight), time in specific intensities, and estimated PAEE | Summary graphs and reports by day | PC software, Web-based interface with social networking functions | CCLB | 1 yr | 45 d | 59 + 7 (monthly fees for data tracking) |
| Gruve (www.gruve.com) | Omnidirectional monitor | No raw data accessible: summary measures include estimated TDEE | 60 s epochs data by day and graphics by day, hour, or minute | Web-based data server with personal summaries | RLIB | 3–4 d (2–3 h to charge) | 20 | 99 + 14.95 (monthly fees) |
| Omron HJ-720ITC GoSmart Pedometer (www.omronhealthcare.com) | Biaxial accelerometer | Provides reports of steps, aerobic steps and minutes, and distance | 60 s epochs data by hour and by day and graphics by day | PC software with personal summaries | CCLB | 6 months | 42 d | 65 |

Monitors summarized here are provided as examples of different types of monitoring technologies. The list should not be assumed to be comprehensive or definitive. For additional information, contact the manufacturers.
RLIB, rechargeable lithium ion battery; CCLB, coin cell lithium ion battery; PS, power supply; BL, battery life; MEM, memory; PAEE, PA energy expenditure; TDEE, total daily energy expenditure; GSR, galvanic skin response; PA, physical activity.

systematically if we adopted a similar approach with accelerometer development and validation.

A monitor equivalency study can be viewed as a form of cross-validation study in which the validity of a monitor is compared with a criterion measure or with other alternative monitors. If multiple monitors are compared with the same criterion, then the relative validity of each monitor is directly compared with the criterion measure (criterion validity). If monitors are compared with each other to test agreement, then the study is focused more on convergent (or concurrent) validity. Criterion evidence is always desirable, but knowing whether two instruments provide equivalent information can still be helpful. Recommendations for best practices are drawn primarily from measurement literature. Measurement theorists emphasize that establishing validity is a process and that because multiple pieces of information are needed to document validity, it cannot be accomplished in a single study. The following guidelines are provided to standardize and improve the quality of monitor comparison studies. The focus is on validity comparisons rather than reliability comparisons because validation studies are more common and the more pressing need. Reliability also is a prerequisite for validity, so major problems with reliability will be evidenced by weak validity evidence.

## Outcome Considerations

Because many competing accelerometry-based devices are now available, it is important to compare them on common grounds. The most appropriate outcome measure is energy expenditure because this outcome can be directly related to levels of physical activity using accepted MET-based categorization (e.g., moderate-intensity activity between 3 and 5.9 METs) (3,4). The other major advantage of standardizing on the basis of energy expenditure is that methods are available to provide appropriate criterion data, such as indirect calorimetry or doubly labeled water (16).

Standardizing outcome measures on the basis of energy expenditure is critical because it enables different technologies and approaches to be directly compared. An example from another commonly used monitoring device illustrates this point. Automated sphygmomanometers measure blood pressure, and precision likely varies on the basis of the quality of the blood pressure cuff as well as the precision of the gauge. However, the outcome (e.g., mm Hg) is at least standardized and directly comparable in all blood pressure cuffs. By standardizing outcomes for comparisons on the basis of units of energy expenditure (or METs), it will be possible to directly compare accuracy of and agreement between different devices.

That said, energy expenditure may not be an appropriate criterion measure for all instruments because some instruments are also designed to provide behavioral outputs for time spent in specific postures (e.g., lying, sitting, standing) or time spent walking or performing moderate-to-vigorous physical activity. With advances in instrument technology and data processing methods, the options for specific behavioral outputs are likely to increase. Other approaches (e.g., direct observation) may be needed to validate behavioral outputs reported in their raw form (e.g., detected sitting or standing minutes).

## Sample Considerations

Care must be given to recruit a reasonably representative sample population. A target population should be selected to focus the research and to provide a target for recruitment. Discrete groups that typically need to be tested in independent studies include children (younger than 10 yr), adolescents (10 to 20 yr), adults (20 to 60 yr), and older adults (60 yr and older). Comparisons also are typically needed between overweight and normal-weight participants, but it is recommended that these subgroups be included in all sample populations. Studies that assess the reliability or validity of a monitoring device in only overweight or only normal-weight individuals have limited utility because they provide no insight about how variable results would be if the population was more diverse and representative. Both sexes also should be included in most studies if possible to allow direct comparisons.

Sample sizes for validity studies may depend on the tasks performed and the expected variation in output scores (7). Variation may be caused by differences in the efficiency of movement, speed of movement, and sensitivity of the monitor to detect movement performed so enough participants are needed to obtain representative data on the activities. If activities are performed in controlled settings (e.g., walking on a treadmill at a set speed), fewer subjects may be needed within each representative group to compare monitor output. However, if free-living activities are used, movement will be more variable, and larger sample sizes will likely be needed to provide sufficient information on instrument equivalency.

## Design Considerations

Selecting an appropriate protocol is important for calibration and validation research as well as for monitor equivalency research. If the goal is to assess physical activity patterns under free-living conditions, it is important for protocols to reflect common activities for different populations. For example, children have unique activity patterns compared with adults and older adults, so protocols have to be designed to suit the identified target population.

Laboratory-based monitor equivalency studies should incorporate some treadmill-based activities as well as an array of common lifestyle tasks to provide a comprehensive evaluation of the monitor. Including some treadmill and/or walking or jogging activities is important for all studies because most accelerometry-based devices are designed and calibrated specifically to assess locomotor behavior. However, it is important to incorporate other activities to determine the capabilities of assessing typical activities of daily living. For children, activities could include light play, such

as dribbling a ball, coloring, or playing video games. For adults and older adults, lifestyle tasks could include sweeping floors or stacking and carrying objects. If indirect calorimetry is being used as a criterion measure, it is obviously helpful to use a portable metabolic cart to allow evaluation of some nontreadmill activities. At least 4 to 5 min of data should be collected for each activity, and some rest or washout periods can be used between activities. Protocol length can vary but generally should not last more than an hour for most applications because a long period is too burdensome on participants. Field-based studies provide opportunities to compare monitors over extended periods. In these types of designs, it is important to establish clear temporal links between monitors and to obtain some contextual self-report information on what activities are performed.

## Analytical Considerations

Evaluating agreement between monitors is similar to the procedures used in standard calibration and validation research. Guidelines from measurement experts suggest that at least three unique characteristics are needed to demonstrate agreement: 1) the two measures being compared must yield equivalent group estimates (evidenced by nonsignificant differences in the outcome measure), 2) the measures must be associated with each other (evidenced by correlation coefficients), and 3) the measures must be free from bias (evidenced by Bland–Altman plots). Follow-up analyses to examine individual differences also are recommended. The application of these principles for monitor equivalency research is described below.

**Testing group differences.** The key determinant of agreement is whether the monitors provide equivalent information about physical activity or energy expenditure. Emphasis should be placed on total levels of activity or overall estimates of energy expenditure. Many studies have been conducted evaluating the ability of accelerometry-based monitors to assess different activities ("point estimates"), but an evaluation of the total energy expenditure should be the most important outcome. Evaluating agreement in total energy expenditure across all activities and rest breaks provides a comprehensive evaluation of how the monitor would perform under real-world conditions. An evaluation of individual activities can then be performed to determine whether the overall pattern is consistent across activities. It is possible, for example, for monitors to exhibit nonsignificant differences with a criterion (or another monitor) but still differ in estimates for individual activities. By examining overall agreement and point estimates, it is possible to understand the strengths and limitations of different monitors.

If data are normally distributed and two continuous outcome scores are being compared (e.g., criterion-measured and monitor-estimated minutes of moderate-intensity physical activity), then a simple *t*-test provides an appropriate test. With multiple comparisons, an ANOVA is needed. Mixed-model designs are recommended in most cases because they can control for the repeated nature of the data (i.e., participants having data for multiple activities). For categorical data, the chi-square statistic is appropriate for dichotomous outcomes (e.g., meets predetermined criteria, does not meet criteria). The Cohen $\kappa$ statistic may be used to assess intermonitor agreement while taking into account chance agreement. A $\kappa$ statistic of more than 0.40 indicates moderate agreement or better (15).

Sample plots are provided in Figure 2 to highlight supplemental analyses that can further explore the relative accuracy of the monitor(s). Panel A shows a comparison of point estimates for different activities in the protocol. The overall analyses revealed nonsignificant differences in total activity, but the additional analyses revealed that the monitor overestimated some activities and underestimated others. Panel B shows a plot of averaged minute-by-minute data for all individuals. This plot provides additional insights because it shows that the pattern of energy expenditure followed the pattern assessed with indirect calorimetry. In this case, the data demonstrate that energy expenditure dropped during the intervening 2-min rest period between activities. Panel C shows a plot of individual variation in energy expenditure for the difference of the measured versus monitor-estimated energy expenditure. Although overall estimates showed nonsignificant differences, this plot reveals that the monitor overestimated energy expenditure for some individuals but underestimated energy expenditure for others. Investigators are encouraged to conduct this type of supplementary analyses to advance understanding of an error that influences activity assessments.

**Testing overall associations.** Correlation analyses provide an indicator of the overall strength of association between the estimated value and the criterion measure. It is possible for two measures to be correlated but provide different estimates. As a result, significant correlations are not sufficient for evaluating validity or monitor agreement. The Pearson (interclass) correlation coefficient is used for normally distributed data and the Spearman (rank–order) correlation coefficient is used for nonnormally distributed data. The intraclass correlation coefficient often is used with variables that have similar units, as in studies with paired test–retest data or for reliability comparisons. The intraclass correlation coefficient takes into account the variability in a subject's scores between monitors in assessing agreement.

Many factors can influence the magnitude of correlation, including the normality and linearity of data, range of scores, and outliers. Therefore, it is important to enroll participants with wide variation in scores to avoid a clustering of scores. Plotting the data is highly recommended because it provides a way to check for potential outliers and the distribution of the data being correlated. Examining the consistency across subgroups also is important to identify variables such as age, sex, body mass, or fitness levels, which may contribute variability in motion sensor output and measurement agreement. This can be done either by

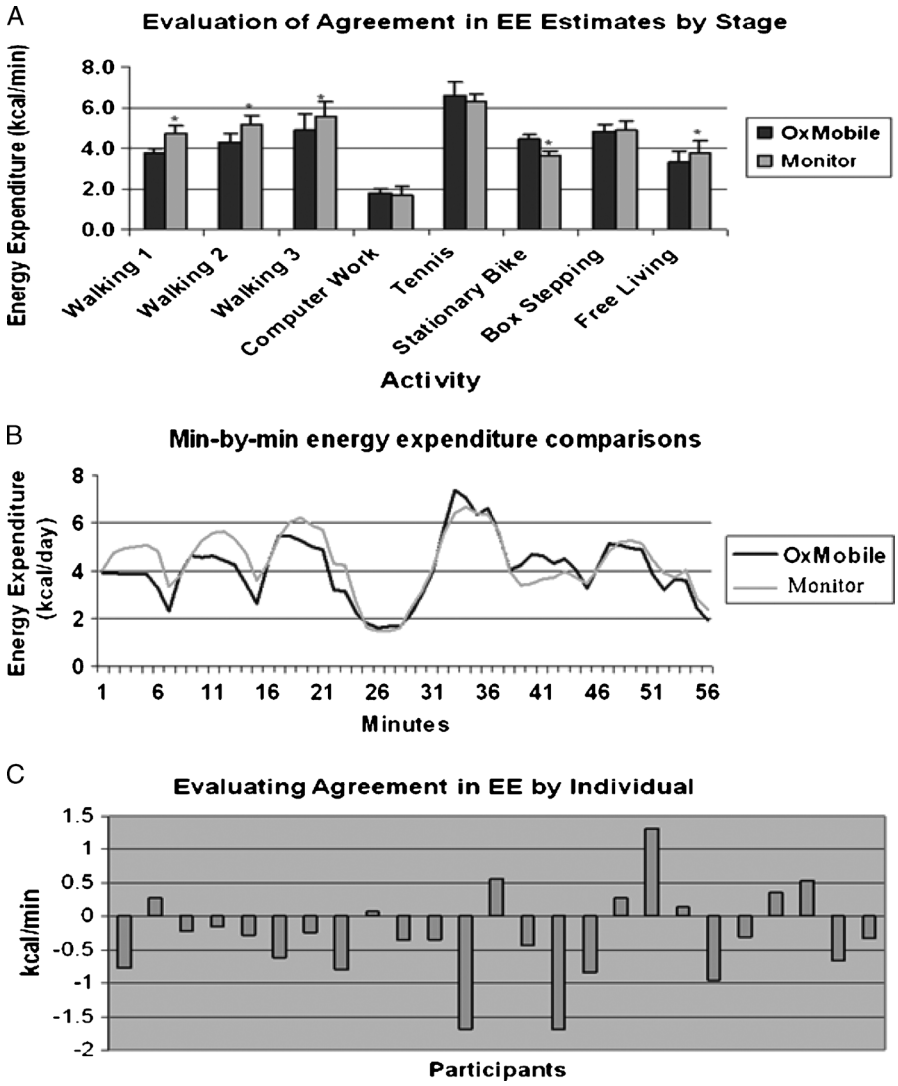


**FIGURE 2—A, Sample image for evaluating agreement with a criterion measure. B, Sample image depicting the creation of average minute-by-minute plots. C, Sample image of individual variability in monitor equivalency studies.**

examining correlations for subgroups or by computing the average individual correlation across multiple individuals. For the latter analyses, correlations are computed for each individual's set of minute-by-minute data, and averages are computed to reflect the overall degree of agreement for the group.

**Testing for potential bias.** The Bland–Altman plot (difference plot) is a method of data plotting to assess agreement between two different variables (5). The Bland–Altman plot has become a standard method to evaluate bias between two scores and is created by plotting the difference between two scores by the mean of the two scores for each subject. By examining the pattern of scores within the means ± 2 SD, investigators can determine whether monitors are consistent in measuring movement across the range of movement scores or whether one monitor is measuring activity differently at higher or lower levels of movement scores.

**Testing for individual differences.** A better understanding of factors that influence individual variability in monitor output or performance is important for advancing research with accelerometry-based activity monitors. Differences in individual agreement in energy expenditure estimation can be examined in further detail using descriptive and correlation analyses. By stratifying participants by some common factor, such as age, height, or weight, it may be possible to find factors that may explain the individual variability. Correlation between factors can reveal the magnitude of the associations. More detailed error analyses also can be conducted to attempt to explain sources of error. Without this additional work, it will be difficult to advance research in this area.

## FUTURE DIRECTIONS

The literature on calibration and validation of monitors has typically demonstrated that monitors are valid for group-level estimation but of limited utility for individual estimation. To allow for individual estimation, we first need to

assess factors influencing individual variability and then address that variability in monitor performance. One way to assess individual variability in monitor performance would be to always directly control for variability in resting metabolic rate. Most studies have assumed the standard MET equivalent of 3.5 $mL \cdot kg^{-1} \cdot min^{-1}$, but this obscures some of the individual variability that exists in metabolic responses. Having better control for differences in body size also is needed to ensure optimal accuracy in assessment.

Advances in monitoring technology and research design will likely contribute to improved accuracy and performance over time. However, it is important for the field to maintain an appropriate balance between accuracy and feasibility. Self-report instruments have typically been the only choice for large-scale research projects where feasibility is a greater concern. This is because they can be administered to large samples of people at a low cost. As the price of accelerometers drop and their ease of use increases, it will become more feasible to use accelerometers for large-scale research applications. However, advances also may detract from utility or keep the price too high for large-scale use. The ability to have data collected during multiple days with a simple noninvasive device may be a more important consideration than obtaining more precise estimates of activity for a given day. The documented utility of pedometers for various field-based applications demonstrates the value of simple measurements for some applications. Individual calibration equations also have been proposed by some investigators, but these procedures add additional costs and may impose logistical constraints that limit more widespread use. Practical low-cost monitoring technologies are still needed.

## CONCLUSIONS

The nature of the research question should be the determining factor in the choice of activity monitor. As illustrated in the vast literature on the health benefits of physical activity (23), powerful effects can be detected with relatively crude measures of activity. Increased precision is clearly needed in some lines of research, but if the goal is to characterize general activity patterns or to determine differences in activity levels between groups, then high levels of precision are probably not necessary. A balance between accuracy and feasibility is clearly needed.

The recommendations provided for conducting monitor equivalency research are intended to help improve the quality of future research with accelerometry-based activity monitors. Although these recommendations do not provide a step-by-step plan or study design (a challenging task given the wide array of instruments and applications for accelerometer-based activity monitors), they provide guidelines that will help advance understanding of the best ways to objectively evaluate measures of physical activity behavior under free-living conditions. Research is needed to better understand our current technology, and it also is important to embrace new technology that may offer direct solutions to these problems. Accelerometry research has evolved considerably during the past 10 yr, and the underlying methodological and analytic work will be put to good use if it can help to inform future research with newer technologies.

The authors thank Pedro Silva and Erik Damen for comments on previous versions of the article as well as Jungmin Lee for evaluation of some newer consumer technologies.

The authors report no conflicts of interest.

## REFERENCES

1. Abel MG, Hannon JC, Sell K, Lillie T, Conlin G, Anderson D. Validation of the Kenz Lifecorder EX and ActiGraph GT1M accelerometers for walking and running in adults. *Appl Physiol Nutr Metab*. 2008;33(6):1155–64.
2. Abraham TL, McClain JJ, Getz RS, Tudor-Locke C. Comparison of low cost objective physical activity assessment instrument versus the ActiGraph. *Med Sci Sports Exerc*. 2008;40(5 suppl):S63.
3. Ainsworth BE, Haskell WL, Leon AS, et al. Compendium of physical activities: classification of energy costs of human physical activities. *Med Sci Sports Exerc*. 1993;25(1):71–80.
4. Ainsworth BE, Haskell WL, Whitt MC, et al. Compendium of physical activities: an update of activity codes and MET intensities. *Med Sci Sports Exerc*. 2000;32(9 suppl):S498–516.
5. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135–60.
6. Brage S, Brage N, Wedderkopp N, Froberg K. Reliability and validity of the Computer Science and Applications accelerometer in a mechanical setting. *Meas Phys Educ Exerc Sci*. 2003;7(2):101–19.
7. Catellier DJ, Muller KE. Sample size and power considerations in physical activity research. In: Welk GJ, editor. *Physical Activity Assessments for Health-Related Research*. Champaign (IL): Human Kinetics; 2002, p. 93–103.
8. Chen KY, Bassett DR Jr. The technology of accelerometry-based activity monitors: current and future. *Med Sci Sports Exerc*. 2005;37(11 suppl):S490–500.
9. Corder K, Brage S, Ramachandran A, Snehalatha C, Wareham N, Ekelund U. Comparison of two ActiGraph models for assessing free-living physical activity in Indian adolescents. *J Sports Sci*. 2007;25(14):1607–11.
10. Esliger DW, Tremblay MS. Technical reliability assessment of three accelerometer models in a mechanical setup. *Med Sci Sports Exerc*. 2006;38(12):2173–81.
11. Grant PM, Ryan CG, Tigbe WW, Granat MH. The validation of a novel activity monitor in the measurement of posture and motion during everyday activities. *Br J Sports Med*. 2006;40(12):992–7.
12. Herrmann S, Hart T, Lee C, Ainsworth B. Evaluation of the Active Key accelerometer. *Br J Sports Med*. 2009;10:1136–40.
13. Jakicic JM, Marcus M, Gallagher KI, et al. Evaluation of the SenseWear Pro Armband to assess energy expenditure during exercise. *Med Sci Sports Exerc*. 2004;36(5):897–904.
14. Krasnoff JB, Kohn MA, Choy FK, Doyle J, Johansen K, Painter PL. Interunit and intraunit reliability of the RT3 triaxial accelerometer. *J Phys Act Health*. 2008;5(4):527–38.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–74.
16. Levine JA. Measurement of energy expenditure. *Public Health Nutr*. 2005;8(7A):1123–32.
17. Macfarlane DJ, Lee CC, Ho EY, Chan KL, Chan D. Convergent validity of six methods to assess physical activity in daily life. *J Appl Physiol*. 2006;101(5):1328–34.

18. McClain JJ, Craig CL, Sisson SB, Tudor-Locke C. Comparison of Lifecorder EX and ActiGraph accelerometers under free-living conditions. *Appl Physiol Nutr Metab*. 2007;32(4):753–61.
18a. McClain JJ, Hart TL, Getz RS, Tudor-Locke C. Convergent validity of 3 low cost motion sensors with the Actigraph accelerometer. *J Phys Act Health*. 2010;7(5):662–70.
19. McClain JJ, Sisson SB, Tudor-Locke C. ActiGraph accelerometer interinstrument reliability during free-living in adults. *Med Sci Sports Exerc*. 2007;39(9):1509–14.
20. McClain JJ, Sisson SB, Washington TL, Craig CL, Tudor-Locke C. Comparison of Kenz Lifecorder EX and ActiGraph accelerometers in 10-yr-old children. *Med Sci Sports Exerc*. 2007;39(4): 630–8.
21. Metcalf BS, Curnow JS, Evans C, Voss LD, Wilkin TJ. Technical reliability of the CSA activity monitor: the EarlyBird Study. *Med Sci Sports Exerc*. 2002;34(9):1533–7.
22. Paul DR, Kramer M, Moshfegh AJ, Baer DJ, Rumpler WV. Comparison of two different physical activity monitors. *BMC Med Res Methodol*. 2007;7:726.
23. Physical Activity Guidelines Advisory Committee. *Physical Activity Guidelines Advisory Committee Report*. Washington (DC): US Department of Health and Human Services; 2008.
24. Powell SM, Jones DI, Rowlands AV. Technical variability of the RT3 accelerometer. *Med Sci Sports Exerc*. 2003;35(10):1773–8.
25. Powell SM, Rowlands AV. Intermonitor variability of the RT3 accelerometer during typical physical activities. *Med Sci Sports Exerc*. 2004;36(2):324–30.
26. Reneman M, Helmus M. Interinstrument reliability of the RT3 accelerometer. *Int J Rehabil Res*. 2010;33(2):178–9.
27. Rothney MP, Apker GA, Song Y, Chen KY. Comparing the performance of three generations of ActiGraph accelerometers. *J Appl Physiol*. 2008;105(4):1091–7.
28. Troiano RP. A timely meeting: objective measurement of physical activity. *Med Sci Sports Exerc*. 2005;37(11 suppl):S487–9.
29. Trost SG, McIver KL, Pate RR. Conducting accelerometer-based activity assessments in field-based research. *Med Sci Sports Exerc*. 2005;37(11 suppl):S531–43.
30. Van Hees VT, Slootmaker SM, De Groot G, Van Mechelen W, Van Lummel RC. Reproducibility of a triaxial seismic accelerometer (DynaPort). *Med Sci Sports Exerc*. 2009;41(4):810–7.
31. Welk GJ, Blair SN, Wood K, Jones S, Thompson RW. A comparative evaluation of three accelerometry-based physical activity monitors. *Med Sci Sports Exerc*. 2000;32(9 suppl):S489–97.
32. Welk GJ, McClain JJ, Eisenmann JC, Wickel EE. Field validation of the MTI ActiGraph and BodyMedia armband monitor using the IDEEA monitor. *Obesity (Silver Spring)*. 2007;15(4): 918–28.
33. Welk GJ, Schaben JA, Morrow JR Jr. Reliability of accelerometry-based activity monitors: a generalizability study. *Med Sci Sports Exerc*. 2004;36(9):1637–45.
34. Zhang K, Pi-Sunyer FX, Boozer CN. Improving energy expenditure estimation for physical activity. *Med Sci Sports Exerc*. 2004; 36(5):883–9.