JOURNAL OF TEACHING IN PHYSICAL EDUCATION, 2004, **23**, 235-251

Measuring Students' Physical Activity Levels: Validating SOFIT for Use With High-School Students

Paul Rowe
Katholieke Universiteit Leuven, Belgium

Hans van der Mars
Oregon State University

Joel Schuldheisz
Concordia University

Susan Fox
Oregon State University

This study was conducted to validate the System for Observing Fitness Instruction Time (SOFIT) for measuring physical activity levels of high-school students. Thirty-five students (21 girls and 14 boys from grades 9-12) completed a standardized protocol including lying, sitting, standing, walking, running, curl-ups, and push-ups. Heart rates and Energy Expenditure, that is, oxygen uptake, served as concurrent validity criteria. Results indicate that SOFIT discriminates accurately among high-school students' sedentary behaviors (i.e., lying down, sitting, standing) and moderate to vigorous physical activity behavior and is recommended for use in research and assessment of physical activity levels in physical education classes for this age group. Implications for use of SOFIT by both researchers and teachers in physical education are described, as well.

Key Words: physical activity measurement, high-school students, assessment.

Introduction

There is compelling scientific evidence documenting the relationship between physical activity and physical and psychosocial health in adults and adolescents (United States Department of Health and Human Services, 1996, 2000).

Rowe is with the Katholieke Universiteit Leuven, Molenstraat 18, Meldert 3320, Belgium. Van der Mars is with the Dept. of Exercise & Sport Science, WB 107C, Oregon State University, Corvallis OR 97331. Schuldheisz is with the Dept. of Physical Education, Concordia University, 2811 NE Holman, Portland, OR 97211. Susan Fox is with Fitness Over 50, Corvallis, OR.

Regular participation in physical activity has been linked with reduced risk for chronic diseases such as vascular and metabolic disturbances, stroke, coronary heart disease, hypertension, obesity, colon cancer, and breast cancer. Regular physical activity can also enhance the immune function and can reduce the risk of upper respiratory tract infections and adult onset diabetes mellitus (e.g., Bouchard & Despres, 1995; Lee, 1995). In addition, it has the potential to postpone or prevent prevalent musculoskeletal disorders and is a key factor for maintaining functional capacity, independence, and a better quality of life (Corbin & Pangrazi, 1993; Vuori, 1995). Finally, physical activity has beneficial effects on anxiety, stress, mild-to-moderate depression, and psychological well-being (e.g., Biddle, 1995; Landers, 1997). These findings underline the public health burden of a sedentary lifestyle and accentuate the health benefits of regular engagement in moderate to vigorous physical activities (Booth & Chakravarthy, 2002). Among behavioral risk factors, physical inactivity, coupled with poor dietary habits, is now considered the second leading cause of death in the United States, behind only smoking (Mokdad, Marks, Stroup, & Gerberding, 2004)

Physical education programs are one of several points of intervention that are part of national efforts to promote physical activity (Heath, 2003). Over the past decade there have been frequent calls for physical education programs to shift their focus from improving the physical fitness of students—a product or outcome orientation—to promoting the physical activity behavior itself —a process orientation (e.g., Corbin, Pangrazi, & Welk, 1994; Freedson & Rowland, 1992). The consensus regarding this shift is most evident in Healthy People 2010 (United States Department of Health and Human Services, 2000) in which the national health objectives relative to physical education are all stated in terms of physical activity behavior. This increased emphasis on physical activity behavior in school-age students has prompted the development of several techniques for assessing physical activity. These techniques include self-report, mechanical devices (such as HR monitoring and motion sensors), the doubly labeled water technique (DLW) (Saris, 1992), and direct observation.

Direct observation systems for measuring physical activity are ideal from a research perspective when behavior is the major focus of concern. Only direct observation allows recording of types, intensity, and duration of physical activity combined with physical environment, social environment, and other contextual characteristics. Improvements have been made in the assessment of physical activity, particularly among children and youth (e.g., McKenzie, 1991; McKenzie, Sallis, & Nader, 1991; McKenzie et al., 1991; Rowe, Schuldheisz, & van der Mars, 1997; Welk, 2002). Other advantages and disadvantages of direct observation for physical activity measurement have been described elsewhere (McKenzie, 1991, 2002).

From a research perspective, valid measurement of students' physical activity levels is a crucial component in the design and evaluation of effective physical education programs; these programs, in turn, have been identified as a primary venue for achieving health-related physical activity goals that extend beyond high school (e.g., Allensworth, Lawson, Nicholson, & Wyche, 1997; Heath, 2003; McGinnis, Kanner, & DeGraw, 1991; Sallis & McKenzie, 1991). McGinnis et al. (1991) noted that "a majority of the national health objectives related to children and youth either directly target school programs and services or can be influenced by the programs that take place in the school setting" (p. 138).

Researchers who use direct observation instruments need assurance that the data collected using such instruments are valid. From a health perspective, content validity is established if and when the behavior categories discriminate among different intensity levels of physical activity. Concurrent validity is optimally demonstrated by a direct measure of energy expenditure (i.e., heart rate or oxygen [O_2] uptake). McKenzie (1991) reviewed eight commonly used direct observation instruments for measuring children's physical activity. Five of the eight instruments included some form of validation through heart rate (HR). HR is relatively easy to assess through reliable and relatively inexpensive HR monitors (Welk, 2002). HR monitors also cause little, if any, physical interference with most activities that students engage in during physical education classes. In contrast to motion sensors, HR monitors provide a practical and physiologically interpretable outcome. Only one instrument, the Children's Activity Rating Scale (CARS) was validated through energy expenditure (EE), as measured by O_2 uptake (Puhl, Greaves, Hoyt, & Baranowski, 1990).

From a theoretical perspective, however, the use of HR as concurrent validity criterion for direct observation instruments also poses a problem. Most validity studies that use HR as a concurrent validity criterion implicitly accept a tight and linear relationship between HR and a direct measure of EE, such as O_2 consumption relative to body weight. This linear relationship is known to be fairly accurate "throughout a large proportion of the aerobic work range." (McArdle, Katch, & Katch, 2001). Nevertheless, for activities involving limited muscle mass and activities at very low or extremely high intensity levels, the relationship between HR and EE has yet to be adequately established. Therefore, the technique of estimating EE from HR has inherent limitations, because HR is really a proxy measure of intensity. On the other hand, O_2 uptake, which is a more direct measure of EE, would logically serve as a stronger concurrent validity criterion, particularly when used for scientific purposes.

The System for Observing Fitness Instruction Time (SOFIT) (McKenzie, 2002; McKenzie, Sallis, & Nader, 1991) is a widely used direct observation instrument. SOFIT includes three coding levels: student physical activity, lesson context, and teacher behavior. The current project targeted only the physical activity level categories that include: lying down (level 1), sitting (2), standing (3), walking (4), and very active (5). Observers record a student's physical activity behavior using momentary time sampling with a 20-s interval length. Thus, every 20 s the observer is cued by a prerecorded audiotape, observes a student's behavior right at that time, and records the appropriate activity level. SOFIT measures the amount of students' physical activity behavior using body position and movement to classify physical activity behavior.

To date, the ages for which most direct observation instruments have been validated range from kindergarten to middle school. Previous validation studies involving SOFIT were carried out with samples from within that age range (McKenzie, Sallis & Armstrong, 1994; McKenzie, Sallis, & Nader, 1991; Rowe et al., 1997). In these instances, HR was used as the criterion in field settings involving simultaneous observation of a standardized protocol of activities.

SOFIT's activity-level categories have not yet been validated for use with high-school-age boys and girls. Several studies and reports have pointed to a gradual decline in physical activity levels through high school, and this decline tends to be more pronounced for female adolescents than for male adolescents (e.g., Centers for Disease Control and Prevention, 2002; Rowland, 1999; Sallis, 1993; United States Department of Health and Human Services, 2000). For example, Sallis (1993) reported declines in physical activity of 2.7% and 7.4% per year for boys and girls, respectively. As a consequence, high school (ages 14-18) is a critical period for promoting continued engagement in physical activity as a lifestyle behavior. Therefore, accurate measurement of physical activity levels and patterns for this age group should be a research priority. The goal for the present study was to replicate previous research by determining the validity of SOFIT's activity-level categories using both HR and O_2 uptake as concurrent validity criteria. For this study a lab setting was chosen in order to maximize standardization of activities across participants and to enable the assessment of O_2 consumption as a direct measurement of EE during activities.

Methods

Participants

Thirty-five students (21 girls and 14 boys), in 9th to 12th grade, from two public high schools volunteered as participants in this study, which was approved by the Institutional Review Board of a university in the western United States. Participant characteristics for both genders are reported in Table 1. The socioeconomic background of the sample ranged from middle to upper-middle class. Both schools were comprised of approximately 90% Euro-American students with the remainder being of Mexican American, African American, Asian American, and Native American descent. This diversity was reflected proportionately in the study's participant sample.

Activity Protocol

Some alterations were made compared with the activity protocols of previous validation studies (e.g., McKenzie et al., 1994; McKenzie, Sallis, & Nader, 1991). The set of activities included tasks reflective of each of the original SOFIT activity-level categories. In addition, two commonly seen muscular strength and endurance tasks (i.e., push-ups and curl-ups) were added to the activity protocol. This protocol was used successfully in a previous field study by Rowe et al. (1997) involving elementary and middle-school students. Applying the activity protocol in a human-performance lab environment allowed for standardization of activities and intensity levels. Students completed the following 42-minute protocol of activities during late afternoons, immediately after leaving school:

1. SOFIT level 1: Lie down on one's back on a bench for 10 consecutive minutes with no movement.
2. SOFIT level 2: Sit on the bench with straight back and feet hanging down for 4 consecutive minutes with minimal movement.
3. SOFIT level 3: Stand on the treadmill for 4 consecutive minutes with minimal movement.
4. SOFIT level 4: Walk on the treadmill (no elevation) for 4 consecutive minutes at 3 mph.
5. SOFIT level 5: Run on the treadmill (no elevation) for 4 consecutive minutes at 5.5 mph.

6. Walk slowly on the treadmill (no elevation) at 2 mph for 4 minutes for cool-down purposes.
7. Rest by lying down (SOFIT level 1) on the bench for 4 consecutive minutes with no movement.
8. Do curl-ups for 2 minutes, using the Fitnessgram (Cooper Institute for Aerobics Research, 1999) protocol for curl-ups (pace is 1 curl-up per 3 s).
9. Rest by lying down on back for 4 consecutive minutes with minimal movement (SOFIT level 1).
10. Do push-ups for 2 minutes, using the Fitnessgram (Cooper Institute for Aerobics Research, 1999) protocol for push-ups (pace is 1 push-up per 3 s).

Based on previous research using HR as the concurrent validity criterion, tasks 8 and 10 have been coded as SOFIT-activity category 5 (very active). Checking the appropriateness of this coding rule was a secondary goal of the present study. Even though both tasks target muscular strength and endurance, they are activity tasks commonly seen in high-school physical education classes.

Data Collection

Students were encouraged to come to the lab in pairs, and one student was tested at a time. Students had not eaten for at least 4 hr before the experiment began and abstained from involvement in vigorous physical activities on the day of the experiment because of the possible effect on HR, O_2 consumption, and their relationship.

Before the start of the activity protocol, students were interviewed to check on compliance with the “do-not-eat” and “refrain-from-vigorous-physical-activity” requests. In addition, students were asked to describe their habitual involvement in extra-curricular physical activities, such as participation in interscholastic sport programs and city recreational sport programs in terms of types of activities, intensity, and weekly involvement over the last year. This information was used to determine the degree to which participants represented a typical sample of the targeted age group regarding their habitual physical activity and engagement in sports.

After height and weight measurements were taken, a heart rate monitor (Polar Vantage™ XL, Polar CIC Inc., New York) was strapped to the students’ chest by a project assistant of the same gender. The validity and advantages of using HR monitors are described by Welk (2002). The HR monitor was programmed to record and store HR measures at 5 s intervals. Next, participants were asked to lie down and relax while the protocol was explained to them. At this point a timer was started indicating the start of the protocol and allowing synchronization of subsequently gathered HR and oxygen uptake (VO_2) values. Students were notified that they could choose to terminate the experiment at any time simply by making a hand signal.

Next, a face mask with a one-way valve was placed over the participant’s nose and mouth. Expired air was led into a Metabolic Measurement Cart (Sensormedics model 2900, Sensormedics Inc., Yorba Linda, California), programmed to measure and store ventilation (VE), respiratory quotient (RQ), and VO_2 every 20 s. Accuracy of measurement was assured by an automated calibration and verification procedure before each test. After 6 minutes of lying down, data collection was initiated simultaneously for the HR monitor and the metabolic cart. Thus, HR and VO_2 values were obtained at 5 s and 20 s intervals, respectively, for the last 36 minutes of the activity protocol. Immediately after the last activity, HR data and VO_2 values were downloaded and stored in a portable computer.

Data analysis

Primary data control. Individual HR data were scanned visually for missing and impossible values (as a consequence of HR monitor malfunctioning) within each task. Single missing values on 5 s intervals occurred occasionally because of transmission errors. Substitution was made using the mean of the previous and subsequent measure. Unexpectedly low ventilation volumes from the metabolic cart typically indicate air leaks. When such values occurred, the corresponding VO_2 measure for that 20 s interval was dropped from the database.

Instrument Reliability. Assessment of instrument reliability, both for HR and EE, was performed using intraclass correlations. Internal consistency reliability of HR and EE measures was calculated for the seven main activities in the protocol. Cooling-down intervals of walking and resting between main activities were not considered for further statistical analysis because these activities had no direct relevance for the study. In this case, HR at 5 s intervals and EE measures at 20 s intervals from the second to the last minute of the activity were used as repeated measures within activities. For most activities this resulted in 36 repeated HR and 9 repeated EE values, with the exception of curl-ups and push-ups, in which 12 repeated HR and 3 repeated EE-values were analyzed. Along with the intraclass correlation, the standard error of measurement (*SEM*) was calculated.

Individual task-by-task HR and EE. Individual task-by-task HR was obtained by calculating the individual's mean for the given task, excluding data from the first minute of the task. The first minute of each exercise period was eliminated in order to assure a stable data path. In a similar manner, individual task-by-task EE values were obtained. The procedure of cutting out data from the first minute of every activity in a standardized protocol was described by Puhl et al. (1990) and Meijer, Westerterp, Koper, and Ten Hoor (1989). Individuals' mean VO_2 values for the second to the last minute of every activity in the protocol were expressed relative to body weight in order to allow for comparison among participants of different body mass (Bar-Or, 1983). Thus, task-by-task EE values were expressed as O_2 consumption in $ml \cdot kg^{-1} \cdot min^{-1}$.

Descriptive statistics. Means and standard deviations of participants' age, height, weight, and self-reported involvement in extra-curricular sports were calculated for the entire sample and for both genders. Averaged data paths were drawn for both genders, representing the typical effect of the given protocol of activities on HR and EE of high-school boys and girls. Separate means for each gender were calculated for HR and EE values for every 5 s and 20 s interval, respectively. Mean and standard deviation of task-by-task HR and EE values were also calculated for the seven main activities involved in the protocol.

Analysis of variance. HR and EE values were analyzed using a 2×7 (Gender $\times$ Task) between-within Multivariate Analysis of Covariance (MANCOVA) with repeated measures on the last effect. Participants were treated as a single age group because age differences in HR and EE for standardized tasks at high-school ages have been found to be small, compared with larger differences at elementary and middle-school ages (Bar-Or, 1983). In addition, no differences in HR among age groups at higher elementary- and middle-school ages were found for a standardized set of activities in previous research (Rowe et al., 1997). Age and height were included as covariates because statistically significant differences were found between boys and girls for these variables. HR and EE values for the seven main tasks and both covariates (age and height) were examined for univariate and multivariate outliers; univariate and multivariate normality; linearity and homogeneity of variance–covariance matrices; and multicollinearity within both cells (boys and girls) per Tabachnick & Fidell (1996).

Follow-up univariate ANCOVA's were performed on HR and EE values for significant main effects. Scheffé post-hoc tests were conducted to locate the source of any significant differences, and effect size (n^2) and the power of the effects were calculated to determine the meaningfulness and generalizability of effects. The alpha level was set at .05 for all analyses, and was adjusted to .025 for both follow-up ANCOVAs.

Results

Participant Characteristics

Age, height, weight, and involvement in extra-curricular sports for boys, girls, and the entire sample are presented in Table 1. On average, the girls were 9 months older, 3.5 inches shorter, and 8.5 pounds lighter than the boys. Differences between boys and girls were significant for height, $t(33) = 3.77$, $p < .001$, and age, $t(33) = -2.14$, $p < .05$). Because both variables showed significant correlations with HR and EE values for a number of the main tasks, these variables were included as covariates when testing further differences between genders. On average, the boys engaged in almost 9 hr of extra-curricular sports per week, compared with almost 5 hr per week for the girls, with large variance for both subsamples. Engagement in extra-curricular sport was not related to HR and EE values in any of the tasks.

Quality of Data

The percentage of missing HR data was less than 1% for every activity in

Table 1 Means and *SD* for Participant Characteristics

Characteristic	All (*N* = 35) mean (*SD*)	Boys (*n* = 14 mean (*SD*)	Girls (*n* = 21) mean (*SD*)
Age (yr)	15.7 (1.1)	15.3 (0.7)	16.0 (1.2)
Height (in)	67.0 (3.1)	69.1 (3.1)	65.7 (2.2)
Weight (lb)	142.0 (21.4)	147.7 (19.8)	138.2 (22.1)
ECS (hr/week)	6.1 (5.9)	8.8 (7.0)	4.9 (5.1)

Note. ECS = extra-curricular sport.

the protocol, and comparable percentages of data were missing in boys (0.43%) and girls (0.38%). After smoothing, 100% of the task-by-task HR was obtained. The metabolic cart output was monitored constantly during testing, allowing for instant adjustments when suspiciously low ventilation values occurred. In some cases, as a consequence of body movement, the facemask allowed air leaks, mostly during the first minute of the walking section of the protocol. In these instances the facemask was adjusted until normal ventilation values appeared, and the affected 20-s observation was dropped from the database. In all cases these missing 20-s values occurred during the first minute of a given activity, which was not included in the calculation of task-by-task EE values.

Instrument Reliability

Internal consistency reliability estimates of HR and EE values for the entire sample are presented in Table 2. Consistency in HR, measured at 5-s intervals from the second to the last minute of every task, was extremely high. The intraclass correlation ranged from R = .992 for push-ups, to R= .998 for sitting. The *SEM* was less than 1 bpm for most activities, except for curl-ups (1.2 bpm) and push-ups (1.7 bpm). Internal consistency of repeated VO_2 values within tasks ranged from R = .77 for standing to R = .93 for running. The smallest *SEM* was obtained for lying down (0.29 $ml \cdot kg^{-1} \cdot min^{-1}$), the largest *SEM* for push-ups (1.52 $ml \cdot kg^{-1} \cdot min^{-1}$). Based on these results, the HR and EE data were believed to be of good quality.

Comparison of HR Response and EE During Activities

Means (boys vs. girls) of consecutive 5-s HR measures throughout the protocol of activities are displayed in Figure 1. Both HR curves show stable data

Table 2 Internal Consistency Reliability of Heart Rate and Energy Expenditure

	Heart rate (bpm)*		Energy expenditure (ml $O_2 \cdot kg^{-1} \cdot min^{-1}$)**	
Task	*R*	*SEM*	*R*	*SEM*
Lying	0.997	0.54	0.82	0.29
Sitting	0.998	0.61	0.78	0.36
Standing	0.997	0.70	0.77	0.33
Walking	0.997	0.83	0.86	0.67
Jogging	0.995	1.21	0.93	0.85
Curl-ups	0.995	0.94	0.83	0.87
Push-ups	0.992	1.67	0.79	1.52

Note. R = Intraclass correlation coefficients based on one-way ANOVA. *SEM* = standard error of measurement. *Repeated measures are 5 s intervals of HR measure (n = 25, K = 36; except for curl-ups and push-ups where n = 35, K=12). **Repeated measures are 20 s intervals of VO_2 measure (n = 3, K=9; except for curl-ups and push-ups where n = 35, K = 3).

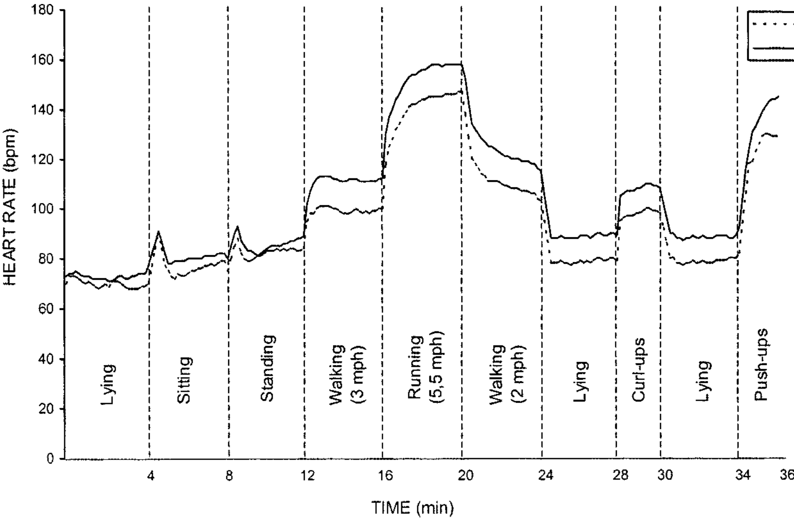

Figure 1 — Mean heart rate (in bpm at 5 s intervals) for boys and girls across tasks.

paths from the second to the last minute of every activity. A small upward trend in HR is evident from lying, to sitting, to standing. Peaks of about 10 bpm occur at the transition between these low intensity activities, demonstrating the sensitivity of the HR monitors when programmed to measure HR at 5-s intervals. For running, a steep initial slope in the first minute levels off, with a small increase in HR remaining (< 10 bpm) from the second to the last minute of this activity. During resting intervals between main activities, HR returned to stable values, somewhat higher (±10 bpm) than the initial lying-down rate. HR during curl-ups was at the same level as during walking, whereas HR during push-ups came up to a level in between walking and running. Although mean HR data are presented, the shapes of both graphs are typical for the individual participants' graphs that were initially used to explore the quality of the data. Although the shape of the data curves for boys and girls are similar, girls had higher HR throughout. Initial gender differences during lying, sitting, and standing were small (< 5 bpm), increasing to a difference of 10 to 20 bpm from the walking section on.

Figure 2 includes the mean EE values (ml.kg^{-1}.min^{-1}) for boys vs. girls. Compared with the HR curve, the EE means did not increase from lying to sitting to standing. Stability in EE values in subsequent activity intervals, from the walking phase of the protocol on, is not as evident as for HR. This explains the lower internal-consistency values reported earlier. Overall, EE values adjusted more slowly to changes in activity intensity than did HR. As with HR, EE values remained slightly higher during rest intervals between main activities than during the initial lying-down interval.

In contrast to HR, EE values during curl-ups remain far below walking values. EE during push-ups was higher than during curl-ups, with the highest values equal to EE during walking. The shape of the curve is representative for the shape of individual data paths. Compared with differences found in HR graphs, differences in EE between boys and girls were inverted: EE means for boys and girls

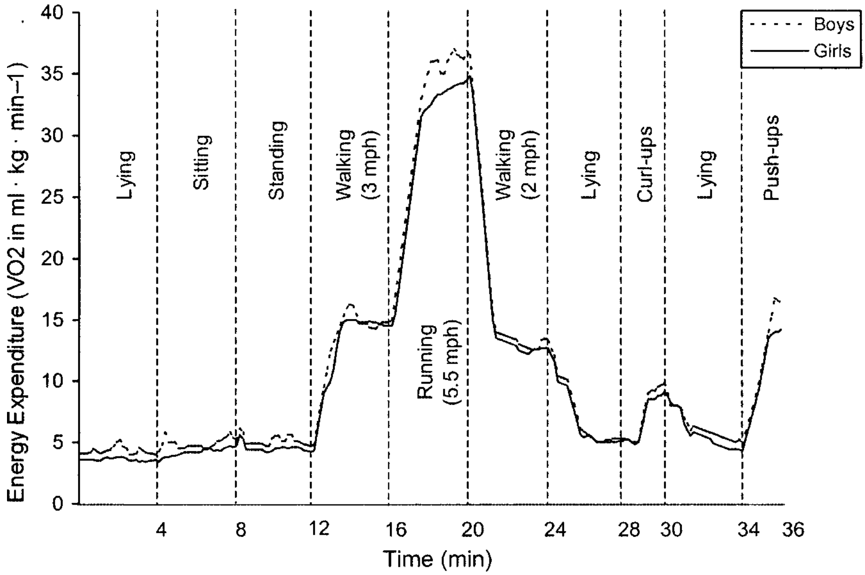

Figure 2 — Mean energy expenditure (in ml $O_2 \cdot kg^{-1} \cdot min^{-1}$ at 20 s intervals) for boys and girls across tasks.

almost overlapped, with slightly lower values for girls, particularly for running, and push-ups.

Descriptive Statistics

Task-by-task HR and EE (means and standard deviations) for main activities are presented in Table 3. As shown in Figure 1, the HR statistics confirm that the pattern of response to activities was consistent across gender, with higher means for girls than for boys. The pattern consists of a small increase in mean HR (± 5 bpm) between lying and sitting, a moderate increase (± 10 bpm) from sitting to standing, a substantial increase (± 20 bpm) from standing to walking, and a large increase (± 50 bpm) from walking to jogging. Mean HR for curl-ups and walking were almost identical, regardless of gender. HR during push-ups was ± 30 bpm higher than during curl-ups and walking.

EE statistics confirm that the pattern of response to activities shown in Figure 2 was consistent across gender, with higher EE means for boys than for girls throughout the protocol. Resting EE values were maintained throughout the first three activities. A strong increase in EE was evident from standing to walking (+11 $ml \cdot kg^{-1} \cdot min^{-1}$) and from walking to jogging (+ 20 $ml \cdot kg^{-1} \cdot min^{-1}$). EE values during curl-ups were located between resting and walking values. Means for EE during walking and push-ups were almost identical for both boys and girls.

Between-Task Differences

Before analysis, HR and EE values for the seven main tasks and both covariates (age and height) were examined for compliance to the assumptions in

Table 3 Means and (*SD*) for Heart Rate and Energy Expenditure

Task	Heart rate (*SD*) bpm			Energy expenditure (*SD*) ml $O_2 \cdot kg^{-1} \cdot min^{-1}$		
	All	Boys	Girls	All	Boys	Girls
Lying	71.6 (9.9)	69.5 (9.8)	73.0 (10.0)	3.9 (0.7)	4.3 (0.7)	3.8 (0.6)
Sitting	77.1 (13.6)	73.5 (11.1)	79.3 (14.9)	4.2 (0.8)	4.5 (0.7)	3.9 (0.7)
Standing	85.7 (12.8)	83.6 (9.1)	87.1 (14.8)	4.2 (0.7)	4.4 (0.7)	4.2 (0.7)
Walking	106.3 (15.1)	98.1 (10.9)	111.7 (15.3)	15.2 (1.8)	15.3 (1.0)	15.1 (2.1)
Jogging	158.6 (17.2)	145.4 (13.6)	167.5 (13.3)	34.2 (3.2)	36.2 (2.5)	33.7 (3.2)
Curl-ups	103.2 (13.3)	96.4 (8.9)	107.8 (13.9)	8.1 (2.1)	8.7 (1.9)	7.7 (2.2)
Push-ups	138.4 (18.7)	130.8 (20.3)	143.4 (16.0)	14.4 (3.3)	15.9 (3.9)	13.5 (2.5)

using a MANCOVA design. Univariate extreme values were found for six cases in two variables, but none of the absolute *z* scores were larger than three. Therefore, these values were left as such. Mahalanobis's D^2 and Cook's statistic revealed no cases with multivariate outliers. All variables showed insignificant skewness and kurtosis and, thus, were accepted as normally distributed.

Homogeneity of variance–covariance needed to be tested because cell sizes were unequal and less than 20 boys participated. Box's M test was not significant ($p > .001$) and, therefore, homogeneity was accepted. Within-cell scatter-plots for all pairs of dependent variables, covariates, and their combinations were examined and showed linear relationships. Multicollinearity within both cells was checked, and none of the dependent variables were shown to be redundant. Therefore, all task-by-task HR and EE values were used in subsequent analyses, and the obtained *F* statistics were considered robust.

The 2 × 7 (Gender × Task) MANCOVA, with HR and EE as dependent variables and height and age as covariates, revealed significant main effects for gender, $F(2, 30) = 4.01$, $p = .029$, effect size = .211, power = .67; task, $F(12, 22) = 354.56$, $p = .000$, effect size = .995, power = 1.00; and Gender × Task, $F(12, 22) = 4.21$, $p = .002$, effect size = .697, power = .99. The follow-up ANCOVA for HR was not significant for gender ($F[1, 31] = 2.92$, $p = .097$, effect size = .086, power = .38) but was significant for task ($F[1, 31] = 491.01$, $p = .000$, effect size = .937, power = 1.00) and Gender × Task ($F[1, 31] = 5.44$, $p = .000$, effect size = .141, power = .99). Significant HR differences were located with Scheffé post-hoc tests. For the task effect, HR was significantly different for all pairs of tasks except for lying down vs. sitting and curl-ups vs. standing. The meaning of the significant task by gender effect can be seen in Figure 3 (see upper tier). HR for boys and girls did not differ significantly for the first three tasks (lying, sitting, standing), whereas girls had a significantly higher mean HR than boys for walking, jogging, curl-ups, and push-ups.

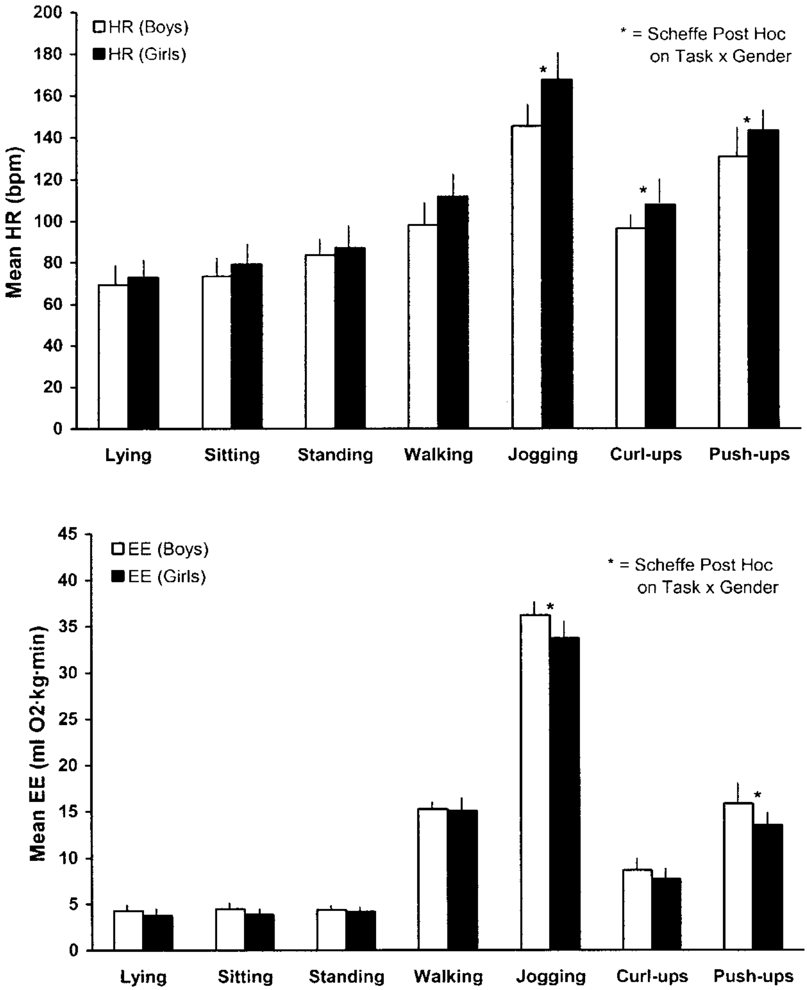

Figure 3 — Mean heart rates (upper tier) and mean energy expenditure (lower tier) for Gender × Task Effect.

The follow-up ANCOVA for EE was not significant for gender (F[1, 31] = 3.68, p = .064, effect size = .106, power = .46) but was significant for task (F[1, 31] = 1427.72, p = .000, effect size = .977, power = 1.00) and Gender × Task (F[1, 31] = 2.90, p = .01, effect size = .08, power = .89). Scheffé post-hoc tests revealed no difference in EE values for lying, sitting, and standing, as well as no difference in EE for walking and push-ups. All other differences were significant. The significant task-by-gender interaction for EE can be interpreted from Figure 3 (see lower tier). Compared with girls, boys had significantly higher EE values for jogging and push-ups.

Discussion

The present study sought to validate the SOFIT systematic observation instrument (McKenzie, Sallis & Nader, 1991) for use with high-school students and to demonstrate the importance of using EE as a concurrent validity criterion. The height and weight for this study's participants were "average" (percentile 50-75) when compared with national anthropometric growth standards (Frisancho, 1990; Najjar & Roland, 1987). Weekly engagement in extra-curricular vigorous physical activities and sports was obtained by self-report. The participants averaged 6 hr of extra-curricular physical activity engagement per week, thus reaching well beyond the physical activity guidelines of 270 minutes of weekly moderate-to-vigorous physical activity (MVPA), particularly because time spent in school physical education classes was not included. Boys engaged in about twice as much extra-curricular physical activity as girls. This finding is along the lines of the gender differences reported by Pate, Long, and Heath (1994), but strongly exceeds the 15–25% difference found across a range of studies. The high rate of engagement in extra-curricular physical activity might have made the sample somewhat select, but this did not correlate with measured heart rates or EE values in the present experiment. Given the nature of the experiment, it is possible that physically active boys and girls were more attracted to volunteer for the study than their less active peers.

The HR response to the activity protocol showed a similarly shaped graph across gender, with higher mean values (+5 to 25 bpm) for girls than for boys. This relative tachycardia in girls, described previously by Bar-Or (1983), is known to range between 10 and 20 bpm and is equally observed in adults, adolescents, and children. When age and height were used as covariates, gender differences in HR were not statistically significant for lying down, sitting, and standing but were statistically significant for all activities at higher intensity levels. The effect size for the significant task-by-gender interaction was small ($n^2 = .141$) and was regarded as practically and theoretically meaningless for the goal of the present study.

The observed EE values for walking and running agree with table-derived values calculated from body weight and running or walking speed (McArdle et al., 2001). Using age and height as covariates, gender differences in EE relative to body weight were small and insignificant for most activities with the exception of running and push-ups. Puhl et al. (1990) observed similar lower EE rates for younger girls at higher activity levels and attributed the differences to better economy of movement by the girls. Though speculative, in the present study the differences likely emerged as a consequence of differences in percentage of body fat. Girls show an increase in body adiposity during adolescence, and, because body fat has a lower metabolism than lean tissue, this results in lower oxygen uptake relative to body mass (Bar-Or, 1983). Differences in the present study would probably disappear if EE was expressed relative to lean body mass.

Implications for Researchers

A comparison of the post-hoc test results for the task effect in both follow-up ANCOVA's for HR versus EE revealed important differences in judging the validity of SOFIT's activity intensity scale. When using HR as the concurrent validity criterion, levels 1 (lying) and 2 (sitting) overlap; the difference between level 2 and level 3 is valid; curl-ups should be coded at the same activity level as walking (level 4); and push-ups refer to a level in between level 4 and 5, significantly different from both. This replicates results from previous research by Rowe et al., (1997).

The interpretation of SOFIT's activity level codes, however, is different when using EE as a validity criterion. The absence of significant differences in EE for lying down, sitting, and standing are in line with results reported by Puhl et al. (1990), who found higher EE in standing than in lying or sitting for 5- to 6-year-old children, but only when standing was combined with a second activity (coloring or throwing/catching a nerf ball). Thus, when using EE as a validity criterion, SOFIT activity level categories 1 through 3 can be combined into a single physical activity category (e.g., "No-MVPA"). This would then produce a dichotomous set of physical activity level categories when coupled with the MVPA category of the original SOFIT instrument (i.e., the merged values of the walking and very active categories). Furthermore, EE values for curl-ups in the present study were significantly lower than for walking and push-ups, which were the same. This suggests that the direct observation method overestimates the true intensity level for curl-ups when using HR as the only concurrent validity criterion.

The fundamental purpose of SOFIT is to directly observe and quantify a person's physical activity behavior. SOFIT's development was based, first and foremost, on the need to detect behavior change in children and youth as a consequence of exposure to fitness-related instruction. Thus, when the assessment and/or research focus is explicitly on students' physical activity behavior, the SOFIT instrument will provide valid data based on HR as the concurrent validity criterion. If, however, the explicit purpose of the research is to obtain an estimate of EE with direct observation as the method for measuring MVPA, then frequently used activities in physical education, such as curl-ups, fall below this level. Thus, when observing high-school students, it would be inappropriate to attribute different levels of EE to the behaviors observed under these categories even when differences in HR are found. Though speculative, the same likely would apply to elementary- and middle-school students.

Implications for Teachers in Physical Education

The results of the present study have important implications from the perspective of assessment of students' in-class behavior through the use of direct observation by teachers. Momentary time sampling has been a widely used observation tactic for collecting data on students' behavior in classroom settings. One attractive feature is that it can be employed by teachers while teaching, if the interval length is stretched sufficiently. That is, observation samples could be taken, for example, once every 60 or 90 s. The merging of SOFIT's three lower level physical activity categories into a "no-MVPA" category produces an easier "yes/no" dichotomous coding decision. Potentially, this offers an accurate and practical data-collection tool for assessing students' in-class physical activity behavior by teachers. That is, teachers could select two or three students during a class and assess their MVPA levels, needing to make only a yes/no decision on each of the students once every 60 to 90 s. During the time between the samples, teachers could attend to their instructional function, thus integrating their assessment efforts with their on-going instructional efforts. Further studies are needed, however, to determine the accuracy of MVPA data that are collected at intervals of 60 to 120 s.

Conclusion

The results of this study warrant two conclusions. First, the HR findings in the current study replicate those from previous validation studies of the SOFIT instrument. Thus, properly trained users of SOFIT can be confident that the resulting data are valid when observing high-school-age students. Second, the choice of concurrent validity criterion (HR vs. VO_2) does influence how well SOFIT's categories actually reflect differences in students' physical activity levels. With HR as the criterion, all of SOFIT's physical activity level categories differ, with the exception that there is no difference between lying down and sitting, nor is there a difference between standing and curl-ups. When using EE as the criterion, there is no difference among lying down, sitting, standing, and curl-ups.

References

Allensworth, D., Lawson, E., Nicholson, L., & Wyche, J. (Eds.). (1997). *Schools and health: Our nation's investment*. Washington, DC: Institute of Medicine, National Academy Press.

Bar-Or, O. (1983). *Pediatric sports medicine for the practitioner*. New York: Springer Verlag.

Biddle, S. (1995). Exercise and psychosocial health. *Research Quarterly for Exercise & Sport*, **66**, 292-297.

Booth, F.W., & Chakravarthy, M.V. (2002). Cost and consequences of sedentary living: New battleground for an old enemy. In C.B. Corbin, R.P. Pangrazi, & D. Franks (Eds.), *President's council on physical fitness & sports research digest, Series 3(16)* (pp. 1-8). Washington, DC: President's Council on Physical Fitness & Sports.

Bouchard, C., & Despres, J.P. (1995). Physical activity and health: Atherosclerotic, metabolic, and hypertensive diseases. *Research Quarterly for Exercise & Sport*, **66**, 268-275.

Centers for Disease Control and Prevention (2002). Youth risk behavior survey—United States, 2001. *Morbidity and Mortality Weekly Report,* **51**, 1-68.

Cooper Institute for Aerobics Research (1999). *Fitnessgram—Test administration manual* (2nd Ed.). Champaign, IL: Human Kinetics.

Corbin, C.B., & Pangrazi, R.P. (Eds.). (1993). The health benefits of physical activity. In *Physical activity and fitness research digest, Series 1(1)* (pp. 1-7). Washington, DC: President's Council on Physical Fitness & Sports.

Corbin, C.B., Pangrazi, R.P., & Welk, G.J. (1994). Toward an understanding of appropriate physical activity levels for youth. In C.B. Corbin & R.P. Pangrazi (Eds.), *Physical activity and fitness research digest, Series 1(8),* 1-8. Washington, DC: President's Council on Physical Fitness & Sports.

Freedson, P.S., & Rowland, T.W. (1992). Youth activity versus youth fitness: Let's redirect our efforts. *Research Quarterly for Exercise & Sport,* **63**, 133-136.

Frisancho, A.R. (1990). *Anthropometric standards for the assessment of growth and nutritional status*. Ann Arbor: University of Michigan Press.

Heath, G.W. (2003). Increasing physical activity in communities: What really works? In C.B. Corbin & R.P. Pangrazi (Eds.*), President's council on physical fitness & sport research digest, Series 4(4)* (pp.1-8). Washington, DC: President's Council for Physical Fitness and Sports.

Landers, D.M. (1997). The influence of exercise on mental health. In C.B. Corbin, R.P. Pangrazi, & D. Franks (Eds*.), President's council on physical fitness & sport research digest, Series 2(12)* (pp.1-8). Washington, DC: President's Council for Physical Fitness and Sports.

Lee, I.M. (1995). Exercise and physical health: Cancer and immune function. *Research Quarterly for Exercise & Sport,* **66**, 286-291.

McArdle, W.D., Katch, F.I., & Katch, V.L. (2001). *Exercise physiology: Energy, nutrition, and human performance* (5th ed.). Baltimore, MD: Lippincott, Williams & Wilkins.

McGinnis, J.M., Kanner, L., & DeGraw, C. (1991). Physical education's role in achieving national health objectives. *Research Quarterly for Exercise & Sport,* **62**, 138-142.

McKenzie, T.L. (1991). Observational measures of children's physical activity. *Journal of School Health,* **61**, 224-227.

McKenzie, T.L. (2002). Use of direct observation to assess physical activity. In G. Welk (Ed.), *Physical activity assessments for health-related research* (pp. 179-195). Champaign, IL: Human Kinetics.

McKenzie, T.L., Sallis, J.F., & Armstrong, C.A. (1994). Association between direct observation and accelerometer measures of children's physical activity during physical education and recess. *Medicine and Science in Sport & Exercise,* **26** (Suppl.), 143 (Abstract).

McKenzie, T.L., Sallis, J.F., & Nader, P.R. (1991). SOFIT: System for observing fitness instruction time. *Journal of Teaching in Physical Education,* **11**, 195-205.

McKenzie, T.L., Sallis, J.F., Nader, P.R., Patterson, T.L., Elder, J.P., Berry, C.C., et al. (1991). BEACHES: An observational system for assessing children's eating and physical activity behaviors and associated events. *Journal of Applied Behavior Analysis,* **24**, 141-151.

Meijer, G.A., Westerterp, H.K., Koper, H., & Ten Hoor, F. (1989). Assessment of energy expenditure by recording heart rate and body acceleration. *Medicine and Science in Exercise & Sport,* **21**, 343-347.

Mokdad, A.H., Marks, J.S., Stroup, D.F., & Gerberding, J.L. (2004). Actual causes of death in the United States, 2000. *Journal of the American Medical Association,* **291**, 1238-1245.

Najjar, M.F., & Roland, M. (1987). *Anthropometric reference data and the prevalence of overweight, United States 1976-80*. Hyattsville, MD: United States Department of Health and Human Services.

Pate, R.R., Long, B.J., & Heath, G. (1994). Descriptive epidemiology of physical activity in adolescents. *Pediatric Exercise Science,* **6**, 434-447.

Puhl, J., Greaves, K., Hoyt, M., & Baranowski, T. (1990). Children's activity rating scale (CARS): Description and calibration. *Research Quarterly for Exercise & Sport,* **61**, 26-36.

Rowe, P.J., Shuldheisz, J.M., & van der Mars, H. (1997). Measuring physical activity in physical education: Validation of the SOFIT direct observation instrument for first to eighth-grade students. *Pediatric Exercise Science,* **9**, 136-149.

Rowland, T.W. (1999). Adolescence: A 'risk factor' for physical inactivity. In C. Corbin, & R. Pangrazi (Eds.), *President's council on physical fitness & sports research digest, Series 3*(6) (pp. 1-8). Washington, DC: President's Council on Physical Fitness & Sports.

Sallis, J.F. (1993). Epidemiology of physical activity and fitness in children and adolescents. *Critical Reviews in Food Sciences and Nutrition,* **33**, 403-408.

Sallis, J.F., & McKenzie, T.L. (1991). Physical education's role in public health. *Research Quarterly for Exercise & Sport,* **62**, 124-137.

Saris, W.H.M. (1992). New developments in the assessment of physical activity in children. In J. Coudert & E. Van Praagh (Eds.), *Pediatric work physiology (Children and exercise XVI*) (pp. 107-114). Paris: Masson

Tabachnick, B.G., & Fidell, L.S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

United States Department of Health and Human Services (1996). *Physical activity and health: A report of the Surgeon General*. Atlanta, GA: United States Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion.

United States Department of Health and Human Services (2000). *Healthy people 2010 (Conference edition—two volumes).* Washington, DC: U.S. Government Printing Office. Retrieved February 15, 2004, from http://web.health.gov/healthypeople/Document/tableofcontents.htm.

Vuori, I. (1995). Exercise and physical health: Musculoskeletal health and functional capabilities. *Research Quarterly for Exercise & Sport,* **66**, 276-285.

Welk, G. (Ed.) (2002). *Physical activity assessments for health-related research*. Champaign. IL: Human Kinetics.