Preventive Medicine **31,** S54–S76 (2000)
doi:10.1006/pmed.1999.0542, available online at http://www.idealibrary.com on IDEAL®

Assessment of Physical Activity among Children and Adolescents: A Review and Synthesis[1]

Harold W. Kohl, III, Ph.D.,*,[2] Janet E. Fulton, Ph.D.,† and Carl J. Caspersen, Ph.D., M.P.H.†

*Baylor Sports Medicine Institute, Baylor College of Medicine, Houston, Texas, 77030; and
†Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion,
Division of Nutrition and Physical Activity, Atlanta, Georgia, 30341

Accurate assessment of physical activity in children and adolescents is a challenge. At least six categories of techniques have been used to assess physical activity among children and adolescents, including self-report, electronic or mechanical monitoring, direct observation, indirect calorimetry, doubly labeled water, and direct calorimetry. Each type of technique carries certain strengths and weaknesses. The purpose of this paper is to review and synthesize available evidence on reliability and validity of physical activity assessment techniques used for children and adolescents.

More than 50 papers published between 1971 and 1997 were reviewed for reliability and validity information with children and adolescents ranging in age from 4 to 17 years. In general, the aggregate of the published data suggests a moderate to high test–retest and interinstrument reliability of physical activity assessment although these findings are less consistent among younger children and when the time period between observations in a test–retest assessment was longer than a few days. Results of validity studies are variable, largely due to the use of different validation standards and study designs. A lack of gender and ethnic comparisons was evident in the review. The available data suggest low to moderate validity for self-report and monitoring measures of physical activity.

The choice for use of a particular method of activity assessment among children and adolescents depends largely on the design of the study and the age of the participants. Greater effort toward developing a consistent standard of measure is necessary to make future advancements in improving the accuracy of physical activity assessment among children and adolescents. Moreover, information relevant to girls and ethnic minority children is urgently needed.

Key Words: **validity; reliability; energy expenditure; methods; synthesis; review.**

INTRODUCTION

Physical activity is inversely associated with several health outcomes in adults. Individuals who are more active and fit have lower risks of all-cause mortality, cardiovascular disease, some cancers, diabetes mellitus, and depressive symptoms and may have a lower risk of obesity and osteoporosis as well [*1–4*]. Given the growing wealth of evidence of the health benefits of a physically active life, efforts to increase physical activity among adults have proliferated [*4*]. Moreover, experts advocate promotion of physical activity among children and adolescents for health enhancement and to instill lifelong behavioral patterns that will result in more active and fit adult populations in the future [*5*].

This rationale rests considerably on two fundamental assumptions: first, that there are inherent acute physical and psychological benefits to physical activity among children and adolescents and, second, that physical activity behaviors between childhood and adulthood are correlated and that physically active children are more likely to grow up to be physically active adults, compared with their inactive peers. These active adults will then be healthier by way of a reduced risk of a variety of health conditions. Although the evidence for tracking of physical activity behaviors is tenuous [*6*], most efforts for physical activity promotion among children and adolescents rely on these rationales as their foundation.

The national health policy of promotion of physical activity among children and adolescents, as well as the need to understand patterns of physical activity, implies a need for valid and reliable (repeatable) measures of physical activity. If accurate data cannot be collected, then trends may not be detected and progress toward health goals cannot be measured. In addition to such surveillance needs, research needs also demand a thorough understanding of the reliability and validity of current assessment instruments.

[1] The use of trade names herein is for identification only and does not constitute endorsement by the U.S. Department of Health and Human Services.

[2] Current address and to whom reprint requests should be addressed at International Life Science Institute Center for Health Promotion, 2295 Park Lake Dr., Suite 450, Atlanta, Georgia 30345. Fax: 770-934-7126. E-mail: hkohl@ilsi.org.

0091-7435/00 $35.00

Several techniques have been used to assess physical activity among children and adolescents. These techniques include self-report, direct observation, mechanical or electronic monitoring, direct or indirect calorimetry, and the doubly labeled water method; however, the applications of each of the techniques are quite different. Data derived from self-report collection, for example, have a relatively low level of precision, whereas laboratory methods such as calorimetry and doubly labeled water offer a greater precision. Costs per observation and difficulty in implementation of laboratory methods, however, limit their applicability to large, population-based studies.

Self-report and monitoring devices, on the other hand, are not as difficult to implement and offer a lower cost per observation, but lack the precision of the laboratory-based measures. Consequently, these laboratory techniques have not been used as extensively as the self-report and monitoring methods, making their applicability in various population subgroups unclear.

Because of the differences in precision, the choice of validation standard is a critical aspect of studies of physical activity assessment. Validation efforts must rely on a more precise method as the choice of a criterion against which to measure a test method. For example, any one of five methods (monitoring, direct observation, indirect calorimetry, doubly labeled water, or direct calorimetry) may be used to validate self-report methods. Self-report, due to a relatively lower precision, should not be used as a standard to validate direct observation techniques (for example).

Although other reviews of physical activity assessment among children and adolescents are available [*7–9*], a comprehensive tabulation and synthesis of all methods including their validity and reliability characteristics does not exist. The purpose of this paper is to review and synthesize available literature describing the validity and reliability of different methods of physical activity assessment used among children and adolescents.

All methods of physical activity assessment have been included, and papers reporting a quantitative comparison of validity or reliability analyses were included for review. Multiple studies presented in a single report are reviewed and summarized as separate studies. Similarly, when more than one research question was addressed in a single study, each question was reviewed separately. For the purposes of this paper, a child is defined as any youth aged 5–11 years and an adolescent as any youth aged 12–18 years. Where appropriate data exist, we have included information on children younger than 5 years.

In evaluating this literature, we distinguish between the different constructs connoted by the terms *physical activity* and *energy expenditure*. Physical activity refers to any bodily movement that results in energy expenditure [*10*] and is often measured in distance, time, or arbitrary units. Energy expenditure therefore is a result of physical activity. However, *total* energy expenditure includes aspects other than physical activity, such as resting metabolism and the thermic effect of food, and is usually measured and reported in kilocalories. Thus, studies of assessment tools designed to measure one construct (e.g., physical activity) that are validated by using a tool that assesses another construct (e.g., energy expenditure) may be limited in the amount of agreement or validity that should be expected because of the imprecise matching of the test method and validation criterion.

RELIABILITY

Reliability is the consistency with which a test or an observer measures what is intended to be measured [*11*] and the extent to which the measurements are repeatable [*12*]. In this paper, the term reliability refers to the consistency of scores or measurements. We differentiate between test–retest reliability (the reliability of a score measured two or more times), interinstrument reliability (between two or more instruments), interobserver reliability (between two or more independent observers making a measurement) and intraindividual variability reliability (consistency of the same measure across different units of time, thus providing an indication of behavioral consistency).

A number of factors may influence reliability including the characteristics of the test, the testing situation, the measurement process, the person making the measurement, the mode of statistical estimation, and the subjects or sample of items being measured (in the case of surveys, the sample population) [*11*]. Moreover, an important consideration when evaluating the reliability of measurements of a behavior, such as physical activity, is an understanding not only of the possible error involved with repeated measurements of the same behavior, but also of the error induced by a lack of stability of the behavior of interest itself, that is, when the behavior varies over time. For example, in a test–retest reliability analysis, a key source of error contributing to a less-than-favorable result could be that the assessment method itself cannot provide reproducible results. However, another explanation could be that the behavior being measured (i.e., physical activity) has changed from the first assessment time to the second. Thus, even though the assessment method may be reliable, changes in the behavior could mask that reliability and make interpretation of a test–retest reliability statistic difficult.

A summary of published literature examining reliability of physical activity assessment methodology used among children and adolescents is presented in Tables

1–3. Table entries are in order of age of the study participants and contain information about relevant characteristics of the participants under study, the instrument being tested, design of the study (usually either test–retest or interobserver design, and the time period between observations), and details of analysis and key findings. As stated earlier, when more than one research question was addressed (for example, multiple methods), a separate entry was made in the table, even if the data were contained in the same publication. Where available, details of age, gender, and ethnic distribution of study participants were included.

Direct Observation

Direct observation involves witnessing physical activity behavior while recording it on a coding form or through a handheld computer device [*9*]. Typically this method involves study personnel who observe a specific child, either in real time or on a videotape, for a certain length of time or series of times. These data are then recorded and converted to some type of summary score. Direct observation has been used in both home and school settings. Direct observation techniques, although not practical for large population studies of physical activity because of a relatively high cost per observation, can be useful for smaller methodologic studies. This technique is especially useful for studies of young children who have not yet developed the cognitive ability to accurately recall detailed information [*13*]. Although direct observation can be reactive (affect the behavior being measured) and can be difficult to implement in a large geographic area, the technique can successfully be used in studies in which participants are confined to a defined space (e.g., school playground or gymnasium, home, or practice field). This can be important as well if investigators are interested in the actual context or environment in which the physical activity occurs.

Ten studies [*14–22*] examined the reliability of direct observation of physical activity in children and adolescents (Table 1). As shown in the table, each study approached the issue of reliability using somewhat different designs, making comparison among studies and synthesis problematic. Six of the 10 studies (60%) [*14,15,17,18,21,22*] were designed to assess interobserver variability. The remaining studies [*14,16,19,20*] evaluated test–retest (time-dependent) reliability by comparing values observed on the same participant from at least two consecutive time periods. The time periods under study varied in length from 12 h between observations [*20,21*] to 30 days [*14*]. Studies ranged in size from 14 [*14*] to 192 [*15*] participants, and only 1 study [*16*] provided information on the ethnic distribution of children being observed. Children participating in direct observation studies were generally young; most study participants were under 10 years of age.

The studies evaluating interobserver reliability reported exceptionally good agreement, even between different experimental designs. In most cases, the agreement between observers exceeded 90%. Results from studies evaluating test–retest reliability of direct observation, however, were much more variable, perhaps due to differences in the time between observations among the studies. For example, DuRant et al. [*16*] evaluated test–retest reliability over four consecutive days, while Klesges et al. [*14*] evaluated observations taken one month apart. Because a child's physical activity behavior patterns likely vary from day to day [*14*], high test–retest reliability from direct observation techniques should not be expected.

Monitoring

Mechanical motion sensors, accelerometers, and heart rate monitors are often used for measuring physical activity. Motion sensors and accelerometers are generally worn at the hip and record movement as "counts" of activity or as estimated caloric expenditure [*9*]. Heart rate monitors are worn around the chest and record the participant's heart rate during the period of observation.

Different types of monitoring devices have different modes of action. Heart rate monitors, for example, record heart rates at given intervals, whereas accelerometers measure acceleration in one or more planes. Accelerometers measure motion in a uni-, bi-, or triplanar motion. Early versions (CALTRAC) measure motion counts above a certain threshold of movement and are able to, based on a prespecified regression equation, estimate caloric expenditure. Newer devices (TRITRAC and CSA) are able to store activity patterns in on-board memory systems based on acceleration such that an activity profile can be stored and downloaded for analysis. Heart rate monitors gather and store information about heart rate responses to exercise and physical activity within a prespecified range and can thus be used, or interpreted as, a proxy measure of acute physical activity. Pedometers are of varying type but usually provide some measure of the number of steps taken during a given period of time. Specific details on these monitoring devices, including their design and construction and engineering details, are beyond the scope of this paper and are available in the individual publications. Risks of equipment failure, loss, tampering, and costs are issues that must be dealt with when considering monitoring devices for physical activity measurement.

The reliability of a variety of mechanical and electronic monitoring devices has been investigated for groups of children and adolescents. Devices such as heart rate monitors and accelerometers have the advantage of being more cost-effective than direct observation

TABLE 1

Reliability Studies: Direct Observation

Study	Sample	Instrument	Design	Analysis	Key findings
Klesges et al., 1984 [14]	7 girls and 7 boys (mean age, 34.8 ± 7.7 months; 24 to 48 months)	Fargo Activity Time sampling Survey (FATS) direct observation system	Interobserver reliability using two observers each independently assessing children for 60 minutes	Percentage agreement	Percentage of agreement ranged from 91 to 98%
				Kappa coefficient	Mean kappa coefficient was 0.90
Klesges et al., 1984 [14]	7 girls and 7 boys (mean age, 34.8 ± 7.7 months; 24 to 48 months)	FATS direct observation	Test–retest reliability at 1-month interval	Generalizability analysis	Generalizability coefficient was 0.59, with four occasions of measurement estimated to be optimal
Puhl et al., 1990 [15]	192 children (3 to 4 years)	Children's Activity Rating Scale (CARS)	Interobserver reliability between 389 paired observations during a 12-month period	Percentage agreement	Percentage of agreement between trained observers averaged 84.1% (+/− 10.1%)
DuRant et al., 1993 [16]	123 children (4 to 5 years; black, Hispanic, white)	CARS	Test–retest reliability over 4 consecutive days of observation	Per-hour measurement reliability	Correlation coefficient within 1 day was 0.81, across 2 days was 0.54, across 3 days was 0.69
McKenzie et al., 1991 [17]	17 boys 25 girls (4 to 8 years)	Behaviors of Eating and Activity for Children's Health Evaluation System (BEACHES)	Interobserver reliability of 19 randomly scheduled home observations	Percentage agreement	Mean percentage of agreement was 94%
				Kappa coefficient	Median kappa was 0.91, range was 0.69–1.00
McKenzie et al., 1991 [18]	31 3rd, 4th, and 5th grade physical education classes	System for Observing Fitness Instruction Time (SOFIT)	Interobserver reliability, 31 classes independently coded by two observers	Percentage agreement	Percentage of agreement between trained observers ranged from 88.3 to 91.8%
Rowe et al., 1997 [19]	92 boys 81 girls (1st to 8th grades, mean age = 10.6 ± 2.0 years)	SOFIT	Stability reliability over 2 days of heart rate monitoring in 5 activities, curl-ups, and push-ups	Intraclass correlation coefficient	Intraclass correlation coefficient ranged from 0.82 to 0.91.
Baranowski et al., 1987 [20]	14 girls and 10 boys (3rd to 6th grade)	Direct observation of 2-min intervals during two-, 12-h periods	Day 1 versus day 2 consistency	Contingency table analysis	"Inconsistent" aerobic activity between day 1 and day 2. Children who were active on first day were no more likely than inactive peers to be active on the second day
Bailey et al., 1995 [21]	7 girls (mean age, 8.1 ± 1.1 years) 8 boys (mean age, 8.5 ± 1.6 years)	Direct observation of 3-s intervals over 4-h time blocks for 12-h periods recording posture and intensity of movement	Interobserver reliability using two observers each independently assessing children for 4-h observation periods	Percentage agreement	Mean percentage of agreement during 24-min time blocks was 91%
				Kappa coefficient	Mean kappa coefficient was 0.90
O'Hara et al., 1989 [22]	17, 19, and 21 paired observations for 192 children in three separate environments	Children's Physical Activity Form	Interobserver reliability using two observers of physical activity performed during physical education class	Percentage agreement	Percentage of agreement in three different environments was 96.8, 98.0, and 96.1%

for studies having large numbers of participants although their use could also introduce bias. The monitors could, for example, remind study participants that they are being assessed, causing departures from "usual" behavior.

Nine studies [23–30] evaluated the reliability of various monitoring devices among children and adolescents (Table 2). All but four studies in three references [27,28,30] were designed to evaluate test–retest reliability (behavioral stability). As with the reliability studies of direct observation, the time periods investigated in the studies varied, ranging from day-to-day [25,27,29] to up to 6 months [23,24] between tests. Participants were generally young children; 3–5 year olds were the youngest [23] age group evaluated. In the two studies that simultaneously compared readings among instruments [30], one instrument was placed on each hip during a defined period or exercise protocol and results were compared between the instruments. Three studies [23–25] evaluated reliability properties of a heart rate monitor, three studies [29,30] evaluated the CALTRAC accelerometer, two studies [27,28] examined the newer CSA accelerometer, and one early study [26] evaluated an actometer (a self-winding motion sensor worn on the wrist).

In general, the studies on test–retest reliability have yielded moderate to high associations among observations. The study designs and analyses were sufficiently different to make direct comparisons among studies difficult. The two largest studies [23,24], which were of

TABLE 2

Reliability Studies: Monitoring

Study	Sample	Instrument	Design	Analysis	Key findings
DuRant et al., 1992 [*23*]	82 girls and 77 boys (3 to 5 years) (black, Mexican-American, and white)	Heart rate monitor	Test–retest reliability measured up to 12 times in 1 day and 3–6 months later	Intraclass correlation coefficient	Correlation coefficients for various heart rate response indices (for within 1 day of observation) ranged from 0.81 to 0.85 (n = 61) Correlation coefficients for various heart rate response indices (for multiple subsequent days of observation) ranged from 0.65 to 0.66 (n = 110)
DuRant et al., 1993 [*24*]	66 girls and 60 boys (5 to 7 years) (black, Mexican-American, and white)	Heart rate monitor	Test–retest reliability measured up to 12 times in 1 day and 3–6 months later	Intraclass correlation coefficient	Correlation coefficients for various heart rate response indices (for within 1 day of observation) ranged from 0.75 to 0.92 for various heart rate response indices (n = 73) Correlation coefficients for various heart rate response indices (for multiple subsequent days of observation) ranged from 0.56 to 0.81 (n = 116)
Janz et al., 1992 [*25*]	11 participants (7–15 years)	Heart rate monitor	Test–retest reliability over 2 consecutive days of monitoring	Pearson correlation coefficient	Correlation coefficient between days for total activity was 0.70 and for baseline heart rate values was 0.84
Massey et al., 1971 [*26*]	12 girls and 21 boys (6–15 years) (all mentally retarded, and 33% black)	Actometer	Field assessment of test–retest reliability correlating the average of five 45-minute periods in 1 week with the same periods in the following week	Pearson correlation coefficient	Correlation coefficient between two weekly average scores was 0.80
Janz, 1994 [*27*]	15 girls and 16 boys (7–15 years)	CSA accelerometer	Field assessment of day-to-day stability of measures over 3 days	Pearson correlation coefficient	Correlation coefficients among days for average CSA movement counts per minute were: 0.32 (days 2 and 3), 0.49 (days 1 and 2), and 0.53 (days 1 and 3)
Janz et al., 1995 [*28*]	15 girls and 15 boys (7–15 years)	CSA accelerometer	Field assessment of day-to-day stability of measures over 6 days	ANOVA	No significant difference across means of 6 days for four CSA indices including average movement count
Freedson and Evenson, 1991 [*29*]	17 girls and 13 boys (mean age, 7.0 ± 1.2 years)	CALTRAC	Test–retest reliability over 3 consecutive days.	ANOVA	No significant difference between mean CALTRAC activity counts over 3 days ($P > 0.05$)
				Pearson correlation coefficient	Correlation coefficient between days ranged from 0.38 to 0.79
Sallis et al., 1989 [*30*]	6 girls and 9 boys (mean age, 10.8 years)	CALTRAC	Laboratory assessment of interinstrument reliability, in which one instrument was placed on each hip to collect data during a standardized treadmill exercise test	Pearson correlation coefficient	Correlation coefficient between instruments was 0.89
Sallis et al., 1989 [*30*]	12 participants	CALTRAC	Field assessment of interinstrument reliability, in which one instrument was placed on each hip to collect data throughout the course of a day	Pearson correlation coefficient	Correlation coefficient between instruments was 0.96

heart rate monitors, found substantially higher correlations between multiple observations taken during 1 day of observation than between single observations taken during the same day. Moreover, reliability across 2 days separated by 3 to 6 months showed modest correlations (both for multiple and single measures). These findings may reflect that physical activity among children and adolescents is extremely variable on a day-to-day basis, and expectations for reliability of an assessment instrument over multiple days may need to be tempered accordingly. Data from the two largest studies do not suggest that reliability is higher in studies of older children or adolescents compared with younger children.

Self-Report

Self-reports are the most commonly employed procedures to measure physical activity and can involve recall or diary methods and can be either interviewer-administered or self-administered. Self-report methods are generally relatively inexpensive, quick to administer, unobtrusive, and versatile, and several sources of physical activity information can be obtained from an interview, a questionnaire, or a log. Using a self-report method, study participants are often asked to recall information on physical activity participation during a period in the recent past (e.g., 1 day, 7 days, 1 month) or, alternatively, they may be asked about their usual or "habitual" activity behavior. Major disadvantages of self-report methods include the limitations associated with accuracy of recall and individual interpretations of questionnaires. Self-report methods also are subject to the possibility of misuse among populations or subgroups in which no evidence of validity or reliability is available. That a physical activity assessment questionnaire shows certain validity characteristics in a group of urban-dwelling children does not necessarily indicate the same characteristics for rural children. Moreover, most self-report methods are designed to measure leisure-time physical activity in defined periods, creating an added dilemma during use with children and adolescents because most of a child's physical activity is usually not planned and structured as it is for adults. Thus, existing questionnaires are likely inappropriate for activity which is unstructured (i.e., play time), which is a frequent source of physical activity for children and adolescents. Moreover gender differences in physical activities may exist such that questionnaire choices may not be gender-appropriate in all settings.

A summary of reliability studies of self-report methods conducted with children and adolescents is presented in Table 3. Seventeen studies [*25,29,31–39*] are cited among youth ranging in age from 6 [*29*] to 19 [*36*] years. Seven of the studies (41%) [*25,29,31,32*] were conducted with elementary school children. Sample sizes ranged from 12 [*25*] to 1679 [*35*] youths. Most of the of studies (94%) [*25,29,31–39*] assessed test–retest reliability over a period ranging from 45 min [*38*] to 8 years [*36*]. Characteristics of children and adolescents investigated were inconsistent across the studies. For example, ethnicity of the participants was reported in only 1 study [*34*]. Most reports included both boys and girls as study participants, but gender-specific results were reported and compared in only 4 studies [*36,39*]. Two studies [*29,37*] evaluated the reliability of physical activity. Only 1 study [*38*] evaluated interobserver reliability (0.99), most likely because most self-report instruments do not require direct involvement from trained study personnel (i.e., are not interviewer-administered, but rather are based on self-completion). Test–retest reliability coefficients ranged from 0.20 [*36*] to 0.99 [*38*]. Age was positively associated with the reliability of the instruments; higher coefficients were found for older children and adolescents.

Diaries are infrequently used among children and adolescents, probably because they require rigorous adherence to a daily reporting schedule, which is difficult for youths. Two studies evaluated reliability of this type of tool. Freedson and Evenson [*29*] evaluated the reliability of a physical activity record completed by parents for their child and Bouchard et al. [*37*] evaluated a 3-day record used by older adolescents. The latter work, which involved development and testing of a system to actually estimate energy expenditure, reported a high correlation between assessments (r = 0.96), whereas Freedson and Evenson reported more variable findings, suggesting that studies using parental proxy information may have limited utility.

Among the remainder of the studies, either interviewer-assisted or self-administered questionnaires, the different designs make comparison of findings and synthesis of the information difficult. Most studies were designed to evaluate reliability over a relatively short period (from a few hours to a few weeks), and these studies, in general, yielded modest to high correlations (0.51–0.99). For the 7-day Physical Activity Recall, Sallis et al. [*31*] also demonstrated consistently lower test–retest reliability when 4–6 days passed between assessments than when shorter (2–3 days) intervals occurred, suggesting a time dependency or decay in reliability. Studies of longer duration [*34,36*] may be useful in helping to determine the stability of behaviors, but conclusions about the measurement quality of a specific instrument are obscured by time, maturation, and changes in behavior of study participants.

The performance of these methods as they relate to the age of the children under study is of key interest. Although different methods make it difficult to compare results, data from one study [*39*] allow an interpretation of reliability of instruments as it may relate to age of participants. Among 310 5th, 8th, and 11th graders, a

TABLE 3

Reliability Studies: Self-Report

Study	Sample	Instrument	Design	Analysis	Key findings
Freedson and Evenson, 1991 [29]	17 girls and 13 boys (mean age, 7.0 ± 1.2 years)	Physical activity record kept by parent	Test–retest reliability of physical activity record over 3 consecutive days	Pearson correlation coefficient	Correlation coefficients between minutes of activity ranged from 0.36 to 0.72
Janz et al., 1992 [25]	12 participants (7–15 years)	12-h recall of physical activity	Test–retest reliability, conducted 12 h apart	Pearson correlation coefficient	Correlation coefficient between assessments was 0.56
Sallis et al., 1993 [31]	35 girls and 34 boys (4th grade)	Yesterday Activity Checklist (relative intensity of 20 commonly performed activities done for at least 15 continuous minutes in previous 24 h)	Test–retest reliability conducted 3 days apart	Intraclass correlation coefficient	Correlation coefficient between assessments was 0.60
Sallis et al., 1993 [31]	35 girls and 34 boys (4th grade)	Weekly Activity Sum (relative intensity of 20 commonly performed activities done for at least 15 continuous minutes in previous week)	Test–retest reliability conducted 2 weeks apart	Intraclass correlation coefficient	Correlation coefficient between assessments was 0.51
Sallis et al., 1993 [31]	35 girls and 34 boys (4th grade)	Weekly Activity Checklist (relative intensity of 20 commonly performed activities done for at least 15 continuous minutes in previous week)	Test–retest reliability conducted 2 weeks apart	Intraclass correlation coefficient	Correlation coefficient between assessments was 0.74
Sallis et al., 1993 [31]	35 girls and 34 boys (4th grade)	7-day tally (recall)	Test–retest reliability conducted 2 weeks apart	Intraclass correlation coefficient	Correlation coefficient between assessments was 0.68
Craig et al., 1996 [32]	49 girls (8–11 years)	1-year recall of average number of hours per week 33 physical activities of a 4-MET intensity or higher	Test–retest reliability conducted 2 weeks apart	Pearson correlation coefficient	Correlation coefficient between assessments was 0.70
Godin and Shephard 1984 [33]	698 students (7th to 9th grade)	Self-reported current physical activity habits in three intensity levels (strenuous, moderate, mild)	Test–retest reliability conducted 2 weeks apart	Reliability coefficient (type not specified)	Correlation coefficient between assessments was 0.84
Aaron et al., 1993 [34]	499 girls and 540 boys (12 to 16 years) (24% black, 3% Hispanic or Asian, 73% white)	Past year recall of leisure-time physical activity	Test–retest reliability, conducted 1 year apart	Spearman rank order correlation	Correlation coefficient between assessments was 0.55
Brener et al., 1995 [35]	1679 students (7th to 12th grade)	Youth Risk Behavior Survey (four physical activity questions)	Test–retest reliability 2 weeks apart	Kappa coefficient	Kappa coefficients ranged from 64 to 91%
				Prevalence estimates at time 1 and time 2	No significant difference between prevalence estimates measured at time 1 and time 2
Andersen and Haraldsdottir, 1993 [36]	172 girls and 33 boys (15 to 19 years)	1-year recall of hours of activity in different sports	Test–retest reliability 8 years apart	Pearson correlation coefficient	Correlation coefficient between physical activity measured in 1983 and 1991 was 0.20 for girls and 0.31 for boys
Bouchard et al., 1983 [37]	61 participants (mean age, 14.6 years)	3-day record, recording every 15 minutes on 2 weekdays and 1 weekend day	Test–retest reliability 6–10 days apart	Intraclass correlation coefficient	Correlation coefficient between summary assessments of energy expenditure was 0.96
Weston et al., 1997 [38]	112 randomly selected students (8th to 11th grade; median age, 15 years)	Previous day physical activity recall (PDPAR) to assess energy expenditure from the end of the school day to bedtime	Interobserver reliability of scoring procedure for overall energy expenditure scoring and subcategories of play/recreation and exercise/workout	Pearson correlation coefficient	Correlation coefficient between assessments was 0.99
Weston et al., 1997 [38]	90 students (8th to 11th grade; median age, about 14 years)	PDPAR designed to assess energy expenditure from the end of the school day to bedtime	Test–retest reliability approximately 45 minutes apart	Pearson correlation coefficient	Correlation coefficient between assessments was 0.98

TABLE 3—*Continued*

Study	Sample	Instrument	Design	Analysis	Key findings
Sallis et al., 1993 [39]	70 student volunteers (5th, 8th, and 11th grade)	Godin–Shephard activity survey	Test–retest reliability 2 weeks apart	Pearson correlation coefficient	Correlation coefficient between tests was 0.81, increasing correlation with increasing age. Higher reliability coefficients generally observed in boys versus girls
Sallis et al., 1993 [39]	70 student volunteers (5th, 8th, and 11th grade)	5-choice self-rated activity scale	Test–retest reliability 2 weeks apart	Pearson correlation coefficient	Correlation coefficient between tests was 0.89. No apparent age association. Higher reliability coefficients generally observed in boys versus girls
Sallis et al., 1993 [39]	70 student volunteers (5th, 8th, and 11th grade)	7-day Physical Activity Recall	Test–retest reliability 2 weeks apart	Pearson correlation coefficient	Correlation coefficient between tests was 0.77 for overall kcal/day reliability. Increasing correlation with increasing age. Age by intensity interaction of reliability coefficients was evident. Higher reliability coefficients generally observed in boys versus girls

direct age association was found for two of three self-report instruments. When compared with younger students, higher correlations were found among the older children for assessments taken by the Godin–Shephard survey and a modified 7-day Physical Activity Recall. No apparent age association was observed for self-assessment with a five-point activity scale.

Summary of Reliability Studies

In general, these findings support moderate to high reliability of assessment of physical activity by direct observation and monitoring. Interobserver reliability (for direct observation techniques) and interinstrument reliability (for monitoring devices) appear to be high. For self-report techniques, test–retest reliability appears to be lower for younger children than for older children and adolescents. These observations are reinforced by the series of cross-sectional studies by Sallis et al. [*39*], in which reliability was higher for older participants for two of three instruments tested. Test–retest reliability was generally lower in the studies in which the period of time between the assessments was longer. This finding is not unexpected considering a lack of long-term consistency in many behaviors among children and adolescents.

VALIDITY

Validity is the degree to which an instrument measures what it is intended to measure [*12*]. Validation studies of physical activity assessment methods are usually designed to be either indirect or concurrent studies. Indirect (or construct) validation studies are designed to evaluate the extent to which the measurement in question corresponds to constructs or parameters that are theoretically related to it [*40*]. In physical activity assessment, indirect validation studies are designed to assess the extent to which physical activity is related to a parameter or construct to which it may be related, for example change in a measure of cardiorespiratory fitness as measured by a treadmill exercise test. Although there is a substantial genetic contribution to cardiorespiratory fitness, exercise training studies show that increases in physical activity above a sufficient intensity result in improvements in cardiorespiratory fitness for most people. Thus, this improvement in fitness can be used as an indirect validation criterion for increased physical activity behavior as reported (for example) on a questionnaire.

Indirect validation has frequently been used in studies of physical activity assessment because of a lack of a universally available validation standard against which physical activity assessment methods can be compared [*41*]. Frequently used indirect validation criteria include maximal and submaximal aerobic power, any of several measures of body composition, and other various physical performance measures.

Concurrent validation studies, on the other hand, are designed to measure both the “test” and a presumably more accurate measure of the same parameter as the validation criterion at the same or similar point in time [*40*]. For example, a self-reported measure of physical activity can be concurrently validated against energy expenditure as measured by the doubly labeled water technique. Indirect and concurrent validity studies provide different information, in that, with concurrent validation designs the validation criterion is another more precise measure of the same parameter being assessed. Indirect validation techniques, on the other hand, use a by-product or correlate of the physiologic property or phenomenon under study as a validation criterion.

Relevant published studies designed to evaluate the validity of physical activity assessment methods for use with children and adolescents are listed in Tables 4–7. Table entries are listed roughly by age of the study participants and contain relevant information as to the instrument being validated, the standard that was used to validate the instrument, the design of the study (usually a concurrent validation study or an indirect validation study), type of analysis, and key findings. When more than one question was addressed (for example, multiple methods used in the same study group), a separate entry is provided in each table, even if the data were contained in the same publication. Where available, age, gender, and ethnic distribution of study participants were included.

Physiologic Measures (Doubly Labeled Water)

The most valid and reliable criterion for measurement of energy expenditure is the doubly labeled water technique [*42*]. The technique does not require constant supervision or institutionalization of study participants and allows the individuals to maintain a usual lifestyle. The technique is accurate for estimating energy expenditure, but is not applicable to assessment in large, population-based studies because of cost and inherent participant burden, including logistics related to multiple urine collections and laboratory visits. Protocols can range from 12 to 21 days in duration. Further, a key limitation of the technique is that it only provides information on total energy expenditure and does not allow assessment of the types and patterns of physical activity. Because physical activity is a subset of energy expenditure, the doubly labeled water technique may not be optimal for measurement of the physical activity portion of total energy expenditure.

Goran et al. [*43*] conducted an indirect validation study of the doubly labeled water technique (Table 4). The authors studied 30 children 4 to 6 years of age and quantified associations between total energy expenditure, as measured using a 14-day doubly labeled water protocol, and indirect constructs of body composition, resting energy expenditure, and resting heart rate. Modest to high correlations (0.65–0.86) were reported for all body composition measures as they were related to total energy expenditure, whereas no association was found for resting heart rate. The authors did, however, report a significant positive association between energy expenditure and resting heart rate, after controlling for fat-free mass. Although it is somewhat counterintuitive to use an indirect validation technique to validate a measure such as doubly labeled water, these data do provide important information as to the associations of these various physiologic parameters with the doubly labeled water method. The authors suggested that measures of body composition and resting heart rate could be combined to provide a valid alternative to estimate energy expenditure in larger populations.

Direct Observation

Nine studies [*14,15,17–19,21,22,44,45*], mostly among younger children (age range 2–10 years), validated measures of direct observation techniques (Table 5). Six [*14,17–19,22,44*] were concurrent validation studies, while the remaining three were indirect. The latter used oxygen consumption, dietary intake, or body composition as standards. Largely because of the difficulty and logistics of observing a large number of children for an extended period of time, study sizes are generally small; only five studies [*15,19,22,44,45*] evaluated more than 20 participants.

Five studies [14,17–19,22] used a monitoring device as a validation standard (large scale integrated (LSI) or heart rate monitor) and reported a modest to high correlation (0.64 and 0.90) between the standard and the summary of the observation period. Published studies of the validity of direct observation techniques are difficult to synthesize because of the differing designs and populations. Of the nine studies, only four [*17,18, 19,22*] used the same validation standard. When the nine studies are evaluated individually, each suggests some degree of validity in the method. In particular, the study by Bailey et al. [*21*] used indirect calorimetry to validate observations of 30 activities done in a laboratory setting and reported a very high correlation (0.95) between the activities and oxygen consumption. The activities investigated in this study were selected to closely replicate naturally occurring activities while allowing the investigators to measure metabolic rates. Examples of activities evaluated include sitting, standing, walking, climbing, biking, and running.

Monitoring

Results of studies designed to validate various mechanical and electronic monitors are summarized in Table 6. Eighteen studies [*25,27,29,30,46–55,74*] provided information for three general types of monitors: accelerometers (CALTRAC [*29,30,46–52*], TRITRAC [*50*], CSA [*27*]), heart rate monitors [*25,52,53,74*], and others (LSI activity monitor [*49,54*] and pedometers [*55*]). Table 6 is organized by the type of monitor used in the validation study, and individual studies are entered into each of the three sections roughly by increasing age of participants studied.

Ten (59%) of these validation studies were focussed on accelerometers [*27,29,30,46–52*]. Of these 10, 4 [*46–49*]

TABLE 4

Validity Studies: Physiologic Measures

Study	Sample	Physiologic measure	Validation standard	Design	Analysis	Key findings
Goran et al., 1993 [*43*]	14 girls and 16 boys (4 to 6 years) (Caucasian)	Total energy expenditure from doubly labeled water method	Body composition, resting heart rate, resting energy expenditure	Indirect validation of total energy expenditure by biological variables	Pearson correlation coefficient	Correlation coefficients for total energy expenditure were: 0.86 for fat-free mass, 0.83 for body weight, 0.82 for body surface area, 0.80 for resting energy expenditure, 0.74 for height and 0.65 for fat mass
					Multiple regression	Fat-free mass (74%), resting heart rate (7%), resting energy expenditure (5%) significantly predicted total energy expenditure ($R^2 = 0.86$)

used direct observation (including videotaping) of physical activity as the validation standard, 1 [*51*] quantified energy expenditure using whole-body calorimetry, 1 [*29*] relied on a parental diary, 1 [*52*] used indirect validation methods (laboratory-assessed oxygen uptake), and the remaining 3 [*27,30,50*] used some form of heart rate monitoring. All of these studies were concurrent validation studies in which estimates or indices of physical activity from the monitoring device were correlated with those from the validation standard.

Low to moderate correlations were generally reported for all studies of accelerometers. No association between age of study participants and level of validity was apparent. The two studies in which the highest correlation coefficients ($r = 0.81$ to 0.87) for an accelerometer were reported [*48,51*] evaluated study participants in controlled settings (playroom and a calorimeter). The potential reactivity (i.e., influence of the effect of being measured) of these settings may have limited the opportunities for spontaneous activity or activities usually performed at high or very high intensities. Therefore, the range of types and intensity of physical activities in these studies may have been narrower than among studies where participants were unrestricted by the environment.

The studies [*25,52,53,74*] that used heart rate monitors each used different types of validation standards and statistical analyses. Three [*52,53,74*] of the four were concurrent validation studies, and the other was indirect using peak oxygen uptake test as the standard. Livingstone et al. [*53*] compared energy expenditure estimated from continual heart rate monitoring with that measured by the doubly labeled water technique and found a reasonably high agreement between the two methods. However van den Berg-Emons [*74*], in a study of heart rate monitors among children with spastic cerebral palsy found nearly a 37% variance (−16.9 to 20.0%) between energy expenditure estimated by heart rate monitoring and doubly labeled water estimates.

Thus, most of the data on validity of mechanical and electronic monitoring devices have come from concurrent studies. Those studies reporting the highest validity have been conducted in settings that could be restrictive or reactive, thus potentially limiting the interpretability of the data to free-living populations. Most studies have been conducted using the CALTRAC accelerometer, although the CSA and the TRITRAC devices are more technologically advanced.

Self Report Methods

Self-report methods have been the most frequently validated method of physical activity assessment among children and adolescents. In Table 7, 37 validation studies [*13,25,27,30,32–34,37–39,48,56–67*] are cited, each of which examined self-reported assessment of physical activity among children and adolescents ranging in age from 3 [*56*] to 18 [*38*] years. Only 5 studies [*34,38,56,58,65*] reported ethnicity of the participants. Instruments were self- or interviewer-administered recall of physical activity ranging from 1 day [*30,61–63*] to 1 year [*34*]. Proxy reports of the child's activity taken from parents or teachers were generally used to validate physical activity measures for younger children [*56,58,59*]. For the purposes of this report, proxy reports of physical activity are considered with self-reports although the two are clearly not interchangeable.

Concurrent validation criteria used for energy expenditure were the doubly labeled water [*32*], direct observation [*48,61,65,67*], or monitoring [*25,27,30,38,39, 62–64*]. Indirect validation criteria included performance tests, measurement of oxygen uptake, indices of obesity, and parental physical activity scores. For studies among children aged less than 10 years, analyses revealed insignificant validation coefficients, indicating that most self-report instruments do not measure what they are intended to measure among young

TABLE 5

Validity Studies: Direct Observation

Study	Sample	Instrument	Validation standard	Design	Analysis	Key findings
Klesges et al., 1984 [*14*]	7 girls and 7 boys (mean age, 34.8 ± 7.7 months, 24 to 48 months)	FATS direct observation system	LSI activity monitor	Concurrent validation study correlating a composite index of observed physical activity with LSI readings	Pearson correlation coefficient	Correlation coefficient between index and LSI readings was 0.90
Puhl et al., 1990 [*15*]	12 boys and 13 girls (mean age, 5.6 years)	CARS	Oxygen uptake in specific scale activities as well as maximal effort	Indirect validation analyzing oxygen uptake and heart rate values for activities performed in specified CARS categories	ANOVA	Increasing trend of higher oxygen uptake values and heart rates with increasing category of activity
McKenzie et al., 1991 [*17*]	19 children (4 to 9 years)	BEACHES	Heart rate monitor	Concurrent validation comparing simultaneous measurement of heart rate monitoring and participation in five activity categories of SOFIT	Mean ± standard deviation	Mean heart rates increased with increasing intensity of activity categories, ranging from 99 beats per minute lying down to 153 beats per minute being very active
McKenzie et al., 1991 [*18*]	19 children (4 to 9 years); same study population as McKenzie et al., 1991 [*17*]	SOFIT	Heart rate monitor	Concurrent validation comparing simultaneous measurement of heart rate monitoring and participation in five activity categories of SOFIT	Mean ± standard deviation	Mean heart rates increased with increasing intensity of activity categories, ranging from 99 beats per minute lying down to 153 beats per minute being very active
Rowe et al., 1997 [*19*]	92 boys 81 girls (1st to 8th grades, mean age = 10.6 ± 2.0 years)	SOFIT	Heart rate monitor	Concurrent validation comparing simultaneous measurement of heart rate monitoring and participation in five activity categories of SOFIT, curl-ups, and push-ups	Mean ± standard deviation	Mean heart rates increased with increasing intensity of activity categories, ranging from 87 beats per minute lying down to 182 beats per minute while jogging. Mean heart rate during curl-ups was 120 ± 13.7 and during push-ups was 133 ± 13.5 beats per minute
Bailey et al., 1995 [*21*]	2 boys (8 and 10 years) 2 girls (7 and 8 years) 8 boys (mean age, 8.5 ± 1.6 years) 7 girls (mean age, 8.1 ± 1.1 years)	Direct observation of 3-s intervals over 4-h time blocks for 12-h periods recording posture and intensity of movement	Indirect calorimetry	Indirect validation involving replication of 30 activities in the laboratory while measuring VO_2 and heart rate	Pearson correlation coefficient	Correlation coefficient between energy expenditure, measured by VO_2 and heart rate, of 30 activities was 0.95 The percentage of time spent by children in intense activities was 77% for low, 20% for moderate, and 3% for high
O'Hara et al., 1989 [*22*]	18 girls and 18 boys (8 to 10 years)	Children's Physical Activity Form	Heart rate monitor	Concurrent validation comparing simultaneous measurement of heart rate monitoring and direct observation during physical education class	Pearson correlation coefficient	Average correlation coefficient between heart rate and observed activity points was 0.64, with individual case correlation coefficients ranging from 0.26 to 0.90
Hovell et al., 1978 [*44*]	141 girls and 133 boys (3rd to 6th grade)	5-s interval time-sampling recording system	Observation of 13 female and 31 male adults participating in aerobic activities	Concurrent validation comparing observation of children's physical activity with that of adults	Comparison of mean upper body and lower body activity scores between children and adults	Upper-body scores not significantly different between girls and boys; while lower-body scores were significantly higher among boys. Adults had significantly higher activity scores than children
Corbin and Pletcher, 1968 [*45*]	50 children (mean age, 9.9 years)	Single-frame motion picture record of physical education class	7-day recall of children's diet completed by parent; triceps skinfold measurement	Indirect validation correlating indices of physical activity with skinfold measurement and energy intake	Pearson correlation coefficient	Correlation coefficients for total activity index was 0.52 for skinfold measurement and 0.49 for energy intake

TABLE 6

Validity Studies: Monitoring

Study	Sample	Instrument	Validation standard	Design	Analysis	Key findings
Klesges and Klesges, 1987 [*46*]	13 girls and 15 boys (mean age, 8.8 years)	CALTRAC	Indices of physical activity derived from direct observation	Concurrent validation comparing simultaneous CALTRAC counts and indices of physical activity derived from direct observation during 1 day	Spearman correlation coefficient	Correlation coefficient between index and CALTRAC counts was 0.54; with higher correlations observed for girls, children older than 32.5 months, and overweight children
Mukeshi et al., 1990 [*47*]	9 girls and 11 boys (mean age, 2.9 years) (inner-city residents)	CALTRAC	Kilocalories of expenditure estimated from coding of activities derived from videotaping both indoor and outdoor activities	Concurrent validation comparing simultaneous CALTRAC measurements of energy expenditure and energy expenditure derived from videotapes taken separately for indoor and outdoor time periods for a total of 35 1-h observation periods	Correlation coefficient	Correlation coefficients between observed expenditure and CALTRAC were 0.62 for combined indoor/outdoor data, 0.56 for indoor data only, and 0.48 for outdoor data only. After regression adjustment for age, height, and weight, coefficients were 0.25 for combined indoor/outdoor data; 0.47 for indoor data only, and 0.16 for outdoor data only
Noland et al., 1990 [*48*]	29 boys and 22 girls (mean age, 3.9 years)	CALTRAC	Videotaped time in controlled setting; activity rated using CARS observation system	Concurrent validation	Correlation coefficient	Correlation coefficient between average activity score and CALTRAC reading was 0.86. No differences noted between genders
Klesges et al., 1985 [*49*]	12 girls and 18 boys (mean age, 47.7 ± 12.5 months)	CALTRAC	FATS direct observation system	Concurrent validation	Pearson and Spearman rank correlation coefficients	Pearson correlation coefficient between observation and CALTRAC was 0.39. Spearman rank correlation coefficient between composite behavior (rated by intensity) with CALTRAC was 0.20
Freedson and Evenson, 1991 [*29*]	17 girls and 13 boys (mean age, 7.0 ± 1.2 years)	CALTRAC	Physical activity record of child's activity kept by parent	Concurrent validation	Pearson correlation coefficient	Correlation coefficient between CALTRAC and physical activity record was 0.35
Welk and Corbin, 1995 [*50*]	26 boys (mean age, 9.9 years)	TRITRAC R3D	Activity heart rates determined from heart rate monitoring	Concurrent validation comparing simultaneous TRITRAC and heart rate monitoring data measured over average of 11 h per day for 3-day period	Pearson correlation coefficient	Correlation coefficients between various activity indices and heart rate ranged from 0.44 to 0.60, and between TRITRAC total vector magnitude and activity-related heart rate response was 0.58
Sallis et al., 1989 [*30*]	15 girls and 20 boys (mean age, 10.8 years)	CALTRAC	Mean activity heart rate measured by heart rate monitoring	Concurrent validation comparing simultaneous heart rate and CALTRAC counts measured over average of 10.5 h for 2 days	Pearson correlation coefficient	Correlation coefficient for all data across both days of monitoring was 0.49, with higher coefficients observed for boys, children older than 11 years, and children with BMI < 18.3
Bray et al., 1994 [*51*]	40 girls (mean age, 13.0 ± 1.8 years)	CALTRAC	Energy expenditure measured by whole-body calorimeter	Concurrent validation	Pearson correlation coefficient	Correlation coefficients for energy expenditure from CALTRAC were 0.81 with total energy expenditure, 0.82 with sedentary energy expenditure, and 0.87 with waking energy expenditure
					Bland–Altman technique	Systematic bias noted between CALTRAC and energy expenditure measured in calorimeter with CALTRAC consistently underestimating true value

TABLE 6—*Continued*

Study	Sample	Instrument	Validation standard	Design	Analysis	Key findings
Ballor et al., 1989 [*52*]	20 high school, basketball class students; 10 girls and 10 boys (mean age, 15.2 ± 0.4 years)	CALTRAC	Energy expenditure determined from oxygen uptake measured during 30-min simulated basketball practice session in laboratory	Concurrent validation	ANOVA	Estimates of energy expenditure were heart rate (197 ± 72) > CALTRAC (163 ± 49) > video analysis (123 ± 30). CALTRAC not significantly different from actual energy expenditure
					Pearson correlation coefficient	Correlation coefficients between CALTRAC and heart rate was 0.92, between CALTRAC and video analysis was 0.95, and between video and heart rate was 0.89
Janz, 1994 [*27*]	15 girls and 16 boys (7 to 15 years)	CSA accelerometer	Heart rate monitoring (various indices) across 3 consecutive days	Concurrent validation	Pearson correlation coefficient	Average correlation coefficient between average heart rate and CSA average movements (across 3 days) was 0.57, with higher correlation coefficient (0.64) between average heart rate and average time spent above threshold
Van den Berg-Emons, et al. [*74*]	4 girls and 5 boys with spastic cerebral palsy (8 to 13 years)	Heart rate monitor	Total energy expenditure, determined by doubly labeled water method	Concurrent validation	Spearman correlation coefficient	Correlation coefficient between averaged heart rate and total energy expenditure was 0.88
					Bland–Altman technique	Individual estimates of energy expenditure by heart rate method ranged from 16.9% lower to 20.0% higher than doubly labeled water estimates.
Livingstone et al., 1992 [*53*]	17 girls and 19 boys (7 to 15 years)	Heart rate monitor	Total energy expenditure, determined by doubly labeled water method	Concurrent validation	Bland–Altman technique	95% confidence interval for bias was −0.56–+0.01 MJ/day; limits of agreement (mean difference in total energy expenditure ± 2 SD) were −1.99 to +1.44 MJ/day
					Percentage difference	Mean percentage of difference between doubly labeled water and total energy expenditure (MJ/day) ranged from −9.2 to +3.5%
Janz et al., 1992 [*25*]	40 girls (mean age, 10.8 ± 2.8 years)	Heart rate monitor	Physical fitness measured by maximal exercise test on cycle ergometer	Indirect validation study correlating heart rate monitoring over 12-h period with physical fitness	Pearson correlation coefficient	For girls, the correlation coefficient between average net heart rate: with VO_2 peak was 0.36, with 60% heart rate reserve was −0.02, with % fat was −0.31, and with 60% heart rate reserve was −0.09
	36 boys (mean age, 11.0 ± 2.9 years)					For boys, the correlation coefficient of average net heart rate with VO_2 peak was 0.06, with 60% heart rate reserve was −0.10, with % fat was −0.09, and with 60% heart rate reserve was −0.18
Ballor et al., 1989 [*52*]	20 high school students in basketball class; 10 girls and 10 boys (mean age, 15.2 ± 0.4 years)	Heart rate monitor	Energy expenditure determined from oxygen uptake measured during 30-min simulated basketball practice session videotaped in laboratory	Concurrent validation study	ANOVA	Estimates of energy expenditure were heart rate (197 ± 72) > CALTRAC (163 ± 49) > video analysis (123 ± 30), $P < 0.05$. CALTRAC not significantly different from actual energy expenditure ($P < 0.05$)
					Pearson correlation coefficient	Correlation coefficient between CALTRAC and heart rate was 0.92, between CALTRAC and video analysis was 0.95, and between video and heart rate was 0.89

TABLE 6—*Continued*

Study	Sample	Instrument	Validation standard	Design	Analysis	Key findings
Klesges et al., 1985 [*49*]	12 girls and 18 boys (mean age, 47.7 ± 12.5 months)	LSI activity monitor	FATS direct observation	Concurrent validation study	Pearson and Spearman rank correlation coefficients	Pearson correlation coefficient between observation and LSI was 0.38. Spearman rank correlation coefficient between composite behavior (rated by intensity) with CALTRAC was 0.36
LaPorte et al., 1982 [*54*]	22 boys (12 to 14 years)	LSI activity monitor	Energy expenditure estimated by Minnesota Leisure Time Physical Activity Questionnaire	Concurrent validation study comparing activity counts from LSI over a 2-day period with estimated energy expenditure from personal interview using questionnaire.	Correlation coefficient	Correlation coefficient between LSI counts and energy expenditure was 0.02
LaPorte et al., 1982 [*54*]	22 boys (12 to 14 years)	LSI activity monitor	Physical fitness measured by maximal exercise test	Indirect validation comparing activity counts over a 2-day period with two measures of physical fitness: maximal oxygen uptake and duration of symptom-limited test	Correlation coefficient	Correlation coefficients for LSI readings were 0.29 with duration of test, and −0.20 with maximal oxygen uptake values
Saris and Binkhorst, 1977 [*55*]	11 girls and boys (4 to 6 years)	Pedometer	Index of energy expenditure based on observation	Concurrent validation correlating estimated energy expenditure for the most and least active children in a day-care facility (observed for 3 h each morning for 1 week) with pedometer readings	Correlation coefficient	Correlation coefficient between index and pedometer reading was 0.93

children. These findings have led to recommendations that self-reported data not be collected from children less than 10 years of age [*8,68*]. Among children and adolescents 10 years of age or older, when monitoring devices were used as the validation standard, validation coefficients for the self-reported data ranged from 0.03 [*27*] to 0.88 [*38*]; most were in the 0.30–0.50 range. In three studies that used direct observation as a validation criterion [*61,65,67*], various measures of agreement ranged from 46 to 86%, and two other studies [*48*] showed no association.

Only one study [*32*] validated self-reported physical activity against energy expenditure estimated from the doubly labeled water method. The correlation was 0.47 between estimates based on self-report of a 1-year recall of physical activities and estimates from the doubly labeled water method. Despite the questionable design of this study (i.e., a current measure of physical activity being used to validate a recall instrument designed to collect data up to a year in the past), this finding is quite remarkable in that any correlation was found.

The most complete series of studies of the validity of self-report instruments was published by Sallis et al. [*63*] comparing the validity of four different instruments among 66 fourth grade students in a classroom setting. This study was designed to allow comparison of validity among the different formats. The four instruments included two checklists (a 1- and 7-day recall), a weekly activity summary (7-day recall), and a 7-day tally of the number of days a child remembered being physically active. The first three instruments required the children to report activities they remembered doing for more than 15 minutes; each instrument was administered twice. Results were compared against measurements taken with a CALTRAC monitor over a 3-day period.

None of the four instruments was highly correlated with the CALTRAC activity counts. The highest correlation ($r = 0.40$) was observed between the weekly activity summary (a 7-day recall) and the CALTRAC. Correlations were consistently lower for the second administration of each instrument, although the correlations were statistically significant for the checklists at both administrations. The findings were apparently not associated with the length of recall because the correlation coefficients for the 1-day checklist were nearly identical to

TABLE 7

Validity Studies: Self-report

Study	Sample	Instrument	Validation standard	Design	Analysis	Key findings
Bush et al., 1991 [56]	524 4th to 5th grade and 3- to 4 year-old black children	3-day physical activity recall completed by mothers for 3- to 4-year-old child	Mother's 3-day physical activity recall score	Concurrent validation within families, comparing mother's and child's physical activity	Kappa coefficient	Significant agreement ($P <$ 0.0001) found between reports from mother and reports from child
Noland et al., 1990 [48]	21 girls and boys in preschool (mean age, 4.25 years)	Parent rating of activity	Videotape of activity in controlled setting and home observation (n = 8) using CARS observation rating scheme	Concurrent validation	Correlation coefficients	No correlation found between observed activity during 20-minute observation period with at-home reported activity
Noland et al., 1990 [48]	21 girls and boys enrolled in preschool program (mean age, 4.25 years)	Teacher rating of activity	Videotaped time in controlled setting and home observation (n = 8) using CARS observation rating scheme	Concurrent validation	Correlation coefficients	No correlation found between observed activity during 20-minute observation period with at-home reported activity
Bush et al., 1991 [56]	524 4th to 5th grade and 3- to 4-year-old children	3-day recall completed by older child	Mother's 3-day physical activity recall score	Concurrent validation, within families, comparing mother's and child's physical activity	Kappa coefficient	Significant agreement ($P <$ 0.0001) found between mother's and child's physical activity
Huttunen et al., 1986 [57]	31 obese and 31 normal-weight children (5.7 to 16.1 years)	Parental reports of child's physical activity	Comparison of physical activity by obesity status	Indirect validation of differences in reported physical activity in obese and normal-weight children	N/A	Obese children participated less frequently in training teams and had lower grades for sports at school than normal-weight children. No differences noted in frequency of activity or time spent in active pursuits between obese and normal-weight children
Murphy et al., 1988 [58]	89 girls and 124 boys (6 to 18 years) (black and white)	Parental report of global assessment of child's overall activity	Oxygen uptake measured by cycle ergometry	Indirect validation study	ANOVA	Parental reports of greater activity were associated with higher cardiorespiratory fitness of children
Finegan et al., 1991 [59]	125 mothers of 7-year-old children	Play Activity Questionnaire (parent-report measure of children's play preferences)	Mother's reports of their child's activity level	Concurrent validation	Pearson correlation coefficients	Conner's scale correlated 0.01 to 0.42 with factors of Play Activity Questionnaire. Similar findings for Werry–Weiss–Peters scale
Murphy et al., 1990 [60]	92 girls and boys who volunteered in a blood pressure study (10 to 17 years)	Pictorial posters depicting various intensities of activity	Oxygen uptake measured by cycle ergometry	Indirect validation	ANOVA	Children who self-classified into the sedentary category had significantly lower oxygen consumption than that measured among children who self-classified into moderate or vigorous categories
Simons-Morton et al., 1990 [61]	44 students (3rd grade)	Recall of moderate to vigorous physical activity recall	Direct observation	Concurrent validation of physical activity during physical education class	Percentage agreement	Percentage of agreement was 86.3% between observed moderate to vigorous physical activity episodes >10 minutes and those reported by the children
Simons-Morton et al., 1994 [62]	27 3rd graders, 21 5th graders	Physical Activity Interview (PAI)	Heart rate monitoring	Concurrent validation of previous day recall of physical activity using physical activity record as memory aid	Pearson correlation coefficient	Correlation coefficient between minutes ≥180% resting heart rate and PAI-reported moderate- to vigorous-intensity minutes was 0.57 in 3rd graders and 0.72 in 5th graders
Simons-Morton et al., 1994 [62]	27 3rd graders, 21 5th graders	PAI	CALTRAC	Concurrent validation of previous day recall of physical activity using physical activity record as memory aid	Pearson correlation coefficient	Correlation coefficient between CALTRAC counts and PAI-reported moderate- to vigorous-intensity minutes was 0.47 in 3rd graders and 0.63 in 5th graders

TABLE 7—*Continued*

Study	Sample	Instrument	Validation standard	Design	Analysis	Key findings
Sallis et al., 1993 [*63*]	35 girls and 34 boys (4th grade)	Yesterday Activity Checklist (estimate relative intensity of 20 common activities done for at least 15 continuous minutes)	CALTRAC	Concurrent validation of after-school physical activity (single day)	Pearson correlation coefficient	Correlation coefficient between CALTRAC and energy expenditure estimated from the single day checklist was 0.33
Sallis et al., 1993 [*63*]	35 girls and 34 boys (4th grade)	Weekly Activity Sum (relative intensity of 20 common activities done for at least 15 continuous minutes)	CALTRAC	Concurrent validation of after-school physical activity (average of 3 days monitoring)	Pearson correlation coefficient	Correlation coefficient between CALTRAC and energy expenditure estimated over the week was 0.40
Sallis et al., 1993 [*63*]	35 girls and 34 boys (4th grade)	Weekly Activity Checklist (relative intensity of 20 commonly performed activities done for at least 15 continuous minutes)	CALTRAC	Concurrent validation of after-school physical activity (average of 3 days monitoring)	Pearson correlation coefficient	Correlation coefficient between CALTRAC and energy expenditure estimated from weekly recall was 0.34
Sallis et al., 1993 [*63*]	528 children (50% girls) (4th grade)	Yesterday Activity Checklist (relative intensity of 20 common activities done for at least 15 continuous minutes)	CALTRAC	Concurrent validation	Correlation coefficient	Correlation coefficient between CALTRAC and weekday physical activity index score was 0.098 Correlation coefficient between CALTRAC and weekend physical activity index score was 0.093
Sallis et al., 1993 [*63*]	528 children (50% girls) (4th grade)	Parental report of number of days in previous 7 that child had done activities for at least 15 continuous minutes	CALTRAC	Concurrent validation	Correlation coefficient	Correlation coefficient between CALTRAC and physical activity index estimated from parental report was 0.074
Sallis et al., 1993 [*63*]	528 children (50% girls) (4th grade)	Child's report of summer-time organized sports and class participation	CALTRAC	Concurrent validation	Correlation coefficient	Correlation coefficient between CALTRAC and child's report of summer class participation was 0.038 Correlation coefficient between CALTRAC and child's report of participation in summer sports teams was 0.11
Sallis et al., 1996 [*64*]	70 girls 55 boys (5th grade) from four regions of the U.S. (states = CA, LA, MN, TX)	Physical activity checklist interview	CALTRAC	Concurrent validation	Pearson correlation coefficient	Correlation coefficient between CALTRAC and child's report of minutes in moderate to vigorous physical activity was 0.32
			Heart rate monitoring			Correlation coefficient between heart rate monitoring and child's report of minutes in moderate to vigorous physical activity was child's was 0.50

TABLE 7—*Continued*

Study	Sample	Instrument	Validation standard	Design	Analysis	Key findings
Sallis et al., 1996 [*64*]	70 girls 55 boys (5th grade) from four regions of the U.S. (states = CA, LA, MN, TX)	Self-administered physical activity checklist	CALTRAC Heart rate monitoring	Concurrent validation	Pearson correlation coefficient	Correlation coefficient between CALTRAC and child's report of minutes in moderate to vigorous physical activity was 0.30 Correlation coefficient between heart rate monitoring and child's report of minutes in moderate to vigorous physical activity was child's was 0.58
Sallis et al., 1993 [*39*]	93 student volunteers (5th, 8th, and 11th grade)	7-day Physical Activity Recall (recalled "very hard" activity only)	Heart rate monitoring using minutes in heart rate intervals above predetermined intensity thresholds	Concurrent validation	Pearson correlation coefficient	Correlation coefficient between recalled very hard activity and minutes spent in intervals of heart rate >140 bpm was 0.44. There was an increasing correlation with increasing age Correlation coefficients between recalled very hard activity and minutes in intervals of heart rate >160 bpm was 0.53. There was an increasing correlation with increasing age
Sallis et al., 1993 [*63*]	35 girls and 34 boys (4th grade)	7-day recall of after-school physical activity	CALTRAC	Concurrent validation (average of 3 days monitoring)	Pearson correlation coefficient	Correlation coefficient between CALTRAC and energy expenditure estimated from weekly recall was 0.11
Baranowski et al., 1984 [*65*]	24 children (3rd to 6th grade) (38% black, 58% white)	Daily aerobic self-monitoring form using segmented versus whole-day recording approaches	Direct observation over 2 days	Concurrent validation of whole-day versus segmented-day reporting formats were compared against direct observation	Percentage agreement	Average percentage of agreement was 73.4%. Percentage agreement was higher for segmented than whole day format. Percentage agreement for segmented format ranged from 72 to 82%.
Aaron et al., 1993 [*34*]	608 female and 567 males (12 to 16 years) (24% black; 3% Hispanic or Asian, 73% white)	Past year recall of leisure-time physical activity	Fitness estimated from 1-mile walk/run test	Indirect validation	Spearman rank-order correlations	Correlation coefficients between hours per week of activity and fitness were −0.21 for girls, −0.11 for boys, and −0.35 for girls and boys
Weston et al., 1997 [*38*]	48 students (7th to 12th grade)	Previous Day Physical Activity Recall (energy expenditure from the end of the school day through bedtime)	CALTRAC and pedometer counts	Concurrent validation	Pearson correlation coefficient	Correlation coefficients between total relative energy expenditure were 0.88 with CALTRAC and 0.77 with pedometer counts
Weston et al., 1997 [*38*]	12 girls and 14 boys (15 to 18 years) (75% white)	PDPAR to assess energy expenditure from the end of the school day through bedtime	Heart rate monitor estimates of three indices of heart rate response to exercise	Concurrent validation	Pearson correlation coefficient	Correlation coefficient between average % heart rate range during after school period and estimated energy expenditure during same time period was 0.53 Correlation coefficients ranged between 0.37 and 0.63, with a dose-response noted for increasing levels of time spent (while being monitored) over 50% of the heart rate range

TABLE 7—*Continued*

Study	Sample	Instrument	Validation standard	Design	Analysis	Key findings
Craig et al., 1996 [*32*]	45 girls (8 to 11 years)	1-year recall of average number of hours per week in 33 physical activities at a minimum intensity of 4 metabolic equivalents	Nonresting energy expenditure measured by doubly labeled water method	Concurrent validation	Pearson correlation coefficient	Correlation coefficient between physical activity estimates and nonresting energy expenditure was 0.47
Baranowski et al., 1984 [*65*]	9 experimental and 13 control group participants (3rd to 6th grade)	7-day retrospective form	Daily self-monitoring form completed for 7 days	Concurrent validation comparing daily and retrospective self-report instruments during first phase of an intervention study	Pearson correlation coefficients	Correlation coefficients ranged from 0.23 to 0.30
Janz et al., 1992 [*25*]	40 girls (mean age, 10.8 ± 2.8 years) 36 boys (mean age, 11.0 ± 2.9 years)	12-h recall of activity	Heart rate monitor	Concurrent validation	Pearson correlation coefficient	Correlation coefficient between recall of activity and average net heart rate was 0.50, and between recall of activity and minutes >=60% heart rate reserve was −0.38
Janz et al., 1992 [*25*]	40 girls (mean age, 10.8 ± 2.8 years) 36 boys (mean age, 11.0 ± 2.9 years)	Simple activity rating	Heart rate monitor	Concurrent validation	Pearson correlation coefficient	Correlation coefficients between simple activity rating and average net heart rate was 0.35 and between simple activity rating and minutes >=60% heart rate reserve was −0.18
Sallis et al., 1989 [*30*]	15 girls and 20 boys (mean age, 10.8 years)	7-day Physical Activity Recall	CALTRAC	Concurrent validation study comparing CALTRAC counts measured over an average of 10.5 hours for 2 days with self-report physical activity measures from interviews of children at the end of each day	Pearson correlation coefficient	Correlation coefficients between CALTRAC and reports were 0.49 on day 1 and 0.39 on day 2 of monitoring. No overall value reported. Higher correlations observed for girls, children older than 11 years, and children with body mass index (BMI) < 18.3
Sallis et al., 1989 [*30*]	15 girls and 20 boys (mean age, 10.8 years)	7-day Physical Activity Recall	Heart rate monitor	Concurrent validation comparing activity heart rate measured over average of 10.5 h for 2 days with self-report physical activity measures from interviews of children at the end of each day	Pearson correlation coefficient	Correlation coefficients between heart rate and reports were 0.25 on day 1 and 0.52 for day 2 monitoring. No overall value reported. Higher correlations observed for girls, children younger than 11 years, and children with BMI ≥18.3. Findings are inconsistent with CALTRAC findings (above)
Jenner et al., 1992 [*66*]	1,311 Australian boys and girls (mean age, 12 ± 0.4 years)	Number of days per week engaged in exercise for at least 1 h	20-m shuttle run	Indirect validation	Pearson correlation coefficient	Correlation coefficients for physical activity with 20-m shuttle run time were 0.19 for boys and 0.20 for girls
Wallace et al., 1985 [*67*]	11 boys attending summer camp for overweight boys (mean age, 12.5 years)	7-day Physical Activity Recall	Direct observation	Concurrent validation comparing reported and observed physical activity during previous 7 days	Percentage agreement	Overall kcal energy expenditure showed close agreement: 46% agreement on recall of mode of activities performed; 75% agreement on recall of intensity of activities performed
Godin and Shephard 1984 [*33*]	698 students (7th to 9th grade)	Self-reported current physical activity habits at three intensities	Self-reported physical activity scores among competitive swimmers	Concurrent validation study of physical activity instruments (methods not described in report)	Not reported	Scores reported for child and adolescent competitive swimmers differed significantly from scores reported by average children
Janz et al., 1995 [*27*]	15 girls and 15 boys (7 to 15 years)	Activity Rating Instrument	CSA accelerometer	Concurrent validation	Pearson correlation coefficient	Correlation coefficients between self-rating and average movement count ranged from −0.03 to 0.17

TABLE 7—*Continued*

Study	Sample	Instrument	Validation standard	Design	Analysis	Key findings
Janz et al., 1995 [*27*]	15 girls and 15 boys (7 to 15 years)	3-day recall of sweat episodes	CSA accelerometer	Concurrent validation	Pearson correlation coefficient	Correlation coefficients between self-rating and average movement count ranged from 0.46 to 0.48
Janz et al., 1995 [*27*]	15 girls and 15 boys (7 to 15 years)	3-day aerobic activity recall	CSA accelerometer	Concurrent validation	Pearson correlation coefficient	Correlation coefficients between self-rating and average movement count ranged from 0.05 to 0.39
Anderssen et al., 1995 [*13*]	330 girls and 425 boys (mean age, 13.3 ± 0.3 years)	Leisure-time physical activity questionnaire	Parental report	Concurrent validation	Pearson correlation coefficient	Correlation coefficients between parent-reports and child-reports of parental physical activity ranged from 0.42 to 0.56
Bouchard et al., 1983 [*37*]	150 children (mean age 14.6 ± 2.9 years)	3-day record of every 15 minutes of 2 weekdays and 1 weekend day	PWC150, PWC150/kg, sum of six skinfolds, % fat	Indirect validation study of energy expenditure derived from 3-day diary	Intraclass correlation coefficient	Correlation coefficients between mean energy expenditure (kcal/day) were 0.70 with PWC150, 0.27 with PWC150/kg, 0.40 with sum of 6 skinfolds, and 0.30 with % fat.

Note. NA, not applicable. PWC150, physical work capacity extrapolated to heart rate 150 beats per minute.

those for the 7-day checklist. Despite the relatively low correlations, the authors concluded that the checklist format was appropriate for collecting self-reported data from fourth-grade children. However, the authors suggested that because of the relatively low correlation, some combination of monitoring devices and self-report may be a prudent step to increase validity.

Three studies, among children and adolescents aged 7–15 years [*27*], were conducted using the relatively new CSA accelerometer as a validation criterion. The CSA is considered a new generation of accelerometer that has several advantages over other accelerometers, including a smaller size, easier attachment to study participants, and a longer monitoring period. In these studies, however, low to modest correlations ($r = 0.03$ to 0.48) were again found between the questionnaire measures and CSA indices. These results support the low validity findings from other studies of self-report instruments and suggest that the CSA does not correlate more highly with self-reported measures of physical activity than do those from other monitoring instruments.

Summary of Validation Studies

Interpretation of validity data from physical activity assessment techniques is less straightforward than interpretation of data from reliability studies. Much of the problem lies in the fact that no single or widely available validation criterion exists against which to compare test methods. Measures from calorimetry and doubly labeled water methods are precise measures, but are oftentimes not the appropriate validation criterion for studies in free-living individuals because of their own inherent limitations, including cost and feasibility of the protocol. The data in Tables 4–7 demonstrate the variety of validation standards and study designs that have been used to evaluate assessment methods. Only three studies [*32,51,53*] relied on actual measured energy expenditure from either calorimetry or doubly labeled water as the validation standard. These studies, two which evaluated monitoring devices and one which evaluated a self-report measure, provided no strong evidence of higher validity. Many studies have relied on indirect measures as validation criteria, such as estimates of body composition, aerobic power, or physical performance. Although these factors are related to physical activity, either as a precursor or consequence, they also are influenced by other genetic and environmental factors. Thus use of indirect validation standards may serve to lower the shared variance between the two measures and thus make these measures less useful as validation criteria.

The data suggest low to moderate validity for the self-report and monitoring measures of physical activity assessment. Results are quite variable, due in large part to the lack of consistency among study designs and methods across the investigations. The self-reported method that relies on recall has limited utility among young children, due to cognitive limitations [*68*]. The highest associations between the method being tested and the validation criterion were from studies in which participants were restricted to or within a certain environment while being measured. Although no studies have directly tested this particular scenario, the observation suggests that these physical activity assessment methods may have lower validity in situations where children are allowed to behave more freely.

Limitations

Our understanding of physical activity assessment among children and adolescents has many limitations. A better understanding and control of variability in physical activity assessment methods would help reduce these limitations. Most human behaviors vary between, as well as within, persons. Energy expenditure (including types, patterns, and overall level), for example, varies considerably from day to day in most persons, yet assessment methods rarely include an evaluation of more than 7 days of behavior. Moreover, self-report methods are likely to quantify only gross variations from day to day. Continual monitoring over a long period may place unreasonable burdens on study participants as well as induce the possibility of participant reactivity.

In studies of dietary assessment, attempts have been made to quantify the number of days necessary to accurately measure the true level of intake of a given nutrient or nutrients [*69*]. These efforts require an understanding and quantification of interindividual and intraindividual variation, as well as a reasonable population estimate of the average value of the variable of interest. As with components of dietary intake, the number of days of assessment necessary to estimate true energy expenditure will likely vary with factors such as age and intensity of the activity. Recent attempts to address this issue for younger [*23,24*] and older children [*28*] evaluated the number of days necessary to monitor heart rates to achieve a certain level of reliability. Nevertheless, even a reliable instrument can consistently underestimate or overestimate the truth. Therefore similar attempts must also be made to improve the validity (i.e., to closer approximate true energy expenditure) of physical activity assessment among children and adolescents.

Another limitation concerns the inconsistent approach to statistical analyses of reliability and validity data. The Pearson correlation coefficient, which is designed to measure the degree of linear association between two variables, is the most often used statistical approach. Although it is a robust method, it may not be suitable for all applications, particularly when the association between two measures is nonlinear or when an intraclass correlation may be more appropriate. Reliance solely on the correlation coefficient may result in an incorrect conclusion about the agreement between two methods. Other ways to assess agreement include categorical [*70*] and difference methods [*71*] and other approaches to correlation [*72*]. A complete assessment of the reliability and validity characteristics of a method should include multiple approaches.

Very few studies reviewed for this paper included a description of the ethnic distribution of participants and none evaluated the data by ethnicity. Either this information was not reported in the studies or the study groups were ethnically homogenous. The validity and reliability characteristics of physical activity assessment instruments may differ between ethnic groups and this possibility should be investigated in future studies. Moreover, other characteristics such as gender, health status, socioeconomic status, and family status could affect validity and reliability characteristics and should also be investigated.

SUMMARY

Assessment of physical activity among children and adolescents remains a challenge for investigators. Several assessment methods are available, and the choice for their use depends largely on the study design and the age of participants. Until more work is completed in this area, the correct choice of a measure of physical activity assessment for studies and surveillance among children and adolescents is a difficult decision, and a few general recommendations can be made at this time.

- If assessment among children younger than 10 years of age is the goal, self-report recall methods should not be used.
- Direct observation and mechanical monitoring may be the best method for young children, and interviewer assistance may enhance the validity of recall and reporting among older children and adolescents.
- Electronic monitoring is the best choice for detecting and assessing patterns of physical activity (especially in measures of intensity) over an extended period (several days, for example).
- The doubly labeled water technique, while offering an accurate assessment of total energy expenditure, does not provide an estimate of energy expenditure resulting from physical activity or an evaluation of the intensity of the activity and thus has limited applications in physical activity assessment.
- The applicability of physical activity assessment methods across genders and ethnic groups is unknown given a glaring paucity of information in these subgroups.

The variety of methods and studies makes synthesis and summarization of existing validation research tenuous at best. Reliability and validity studies have been published using a number of different study designs, validation standards, assessment periods, and statistical analyses, among other factors. Each of these issues must be considered by investigators when evaluating the best method for use in a particular population. Many other factors in addition to reliability and validity must also be considered including cost, practicality, and the purpose of a study. Regardless of the measure chosen, investigators must be acutely aware of all sources of variability inherent to the specific method and diligently work to minimize those inputs to maximize validity.

The data suggest that test–retest reliability and interinstrument reliability of physical activity assessments among children and adolescents are generally moderate to high. There is evidence of an age association, especially with self-reports, with the recall of younger children being less reliable. Conversely, low to moderate associations have been found for various validation criteria. This finding was to be expected, considering the variety of methods used to assess activity as well as the lack of a consistent validation criterion.

The range of validity results summarized in this paper suggests that the several methods of assessing physical activity are not measuring identical properties or components. Total physical activity is a function of the type of stimulus (mode of exercise), the intensity at which the stimulus is performed, and the duration of a single episode. Over an extended period, the frequency with which an exercise is performed is also important. For example, one type of electronic monitor may not measure the intensity of physical activity as well as a recall instrument or a diary may, but the monitor may more accurately measure duration. If this is true, researchers may need to use multiple methods to more completely assess all components of physical activity.

FUTURE DIRECTIONS

Greater effort toward developing a standard of measure is necessary in this area. Existing studies have used some measure of energy expenditure (estimated or actually measured kilocalories), an index of counts (as made by accelerometers), time spent in a certain "exposure range" (determined by heart rate monitors), or specially formulated indices of physical activity (from self-report). Sallis and colleagues have demonstrated the problems using existing methods to estimate energy expenditure among children and adolescents from generalized metabolic equivalent (MET) tables developed for adults (73). Clearly, error is involved in all estimates, serving to make comparisons among studies difficult.

The existing literature has many glaring weaknesses, first and foremost perhaps being the paucity of studies among girls (or gender-specific analyses of mixed gender populations) and ethnic minorities. In addition to more validation studies of each of the various methods, in different groups of participants, studies must be designed using appropriate validation standards for the implementation of the measure under study. For example, if total energy expenditure is the important parameter to measure, doubly labeled water would be the best choice for a standard. Similarly, if patterns (including bouts of activity at or above a certain threshold of intensity) are of interest, then one of the newer electronic monitoring devices would be most appropriate.

The development and use of valid and reliable physical activity assessment techniques are particularly important for furthering public health efforts to promote physical activity among children and adolescents. Physical activity promotion is a key part of U.S. health policy, and a crucial part of public health promotion is adequate surveillance. Surveillance systems that use a valid and reliable assessment method allow not only for an accurate baseline assessment, but also for a dependable set of ongoing measurements that can be used to measure secular changes.

ACKNOWLEDGMENTS

We thank Tom Baranowski, Ph.D., William Dietz, M.D., Ph.D., Adele Franks, M.D., James Mendlein, Ph.D., and Michael Pratt, M.D., M.P.H., for constructive comments on an earlier version of the manuscript, Maria P. Alexander, M.P.H., and Barbara Gray for technical and editing assistance, and Lois Corcoran and Yolanda Mallett for assistance in manuscript preparation.

REFERENCES

1. American Heart Association. Statement on exercise: benefits and recommendations for physical activity programs for all Americans: a statement for health professionals by the Committee on Exercise and Cardiac Rehabilitation of the Council on Clinical Cardiology, American Heart Association. Circulation 1992;86: 340–4.
2. Bijnen FC, Caspersen CJ, Mosterd WL. Physical inactivity as a risk factor for coronary heart disease: a WHO and International Society and Federation of Cardiology position statement. Bull World Health Org 1994;72:1–4.
3. Pate RR, Pratt M, Blair SN, Haskell WL, Macera CA, Bouchard C, et al. Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. JAMA 1995;273:402–7.
4. U.S. Department of Health and Human Services. Physical activity and health: a report of the Surgeon General. Atlanta (GA): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, 1996.
5. Sallis JF, Patrick K. Physical activity guidelines for adolescents: consensus statement. Pediatric Exerc Sci 1994;6:302–14.
6. Malina RM. Tracking of physical activity and physical fitness across the lifespan. Res Q Exerc Sport 1996; 67(3 Suppl)S48–57.
7. Heath GW, Pate RR, Pratt M. Measuring physical activity among adolescents. Public Health Rep 1993;108(1 Suppl):42–6.
8. Sallis JF. Self-report measures of children's physical activity. J Sch Health 1991;61:215–9.
9. Pate RR. Physical activity assessment in children and adolescents. Crit Rev Food Sci Nutr 1993;33:321–6.
10. Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public Health Rep 1985;100:126–31.
11. Baumgartner TA, Jackson AS. Measurement for evaluation in physical education. Dubuque (IA): Brown, 1982.
12. Nunnally JC. Psychometric theory. New York: McGraw–Hill, 1967.

13. Anderssen N, Jacobs DR Jr, Aas H, Jakobsen R. Do adolescents and parents report each other's physical activity accurately? Scand J Med Sci Sports 1995;5(5):302–7.

14. Klesges RC, Coates TJ, Moldenhauer LM, Holzer B, Gustavson J, Barnes J. The FATS: an observational system for assessing physical activity in children and associated parent behavior. Behav Assess 1984;6:333–45.

15. Puhl J, Greaves K, Hoyt M, Baranowski T. Children's Activity Rating Scale (CARS): description and calibration. Res Q Exerc Sport 1990;61(1):26–36.

16. DuRant RH, Baranowski T, Puhl J, Rhodes T, Davis H, Greaves KA, et al. Evaluation of the Children's Activity Rating Scale (CARS) in young children. Med Sci Sports Exerc 1993;25: 1415–21.

17. McKenzie TL, Sallis JF, Nader PR, Patterson TL, Elder JP, Berry CC, et al. BEACHES: an observational syste, for assessing children's eating and physical activity behaviors and associated events. J Appl Behav Anal 1991;24:141–51.

18. McKenzie TL, Sallis JF, Nader PR. SOFIT: System for Observing Fitness Instruction Time. J Teach Phys Educ 1991;11:195–205.

19. Rowe PJ, Schuldheisz JM, van der Mars H. Validation of SOFIT for measuring physical activity of first- to eighth-grade students. Pediatr Exerc Sci 1997;9:136–49.

20. Baranowski T, Hooks P, Tsong Y, Cieslik C, Nader PR. Aerobic physical activity among third- to sixth-grade children. J Dev Behav Pediatr 1987;8:203–6.

21. Bailey RC, Olson J, Pepper SL, Porszasz J, Barstow TJ, Cooper DM. The level and tempo of children's physical activities: an observational study. Med Sci Sports Exerc 1995;27(7):1033–41.

22. O'Hara NM, Baranowski T, Simons-Morton BG, Wilson BS, Parcel G. Validity of the observation of children's physical activity. Res Q Exerc Sport 1989;60:42–7.

23. DuRant RH, Baranowski T, Davis H, Thompson WO, Puhl J, Greaves KA, et al. Reliability and variability of heart rate monitoring in 3-, 4-, or 5-yr-old children. Med Sci Sports Exerc 1992;24(2):265–71.

24. DuRant RH, Baranowski T, Davis H, Rhodes T, Thompson WO, Greaves KA, et al. Reliability and variability of indicators of heart-rate monitoring in children. Med Sci Sports Exerc 1993; 25:389–95.

25. Janz KF, Golden JC, Hansen JR, Mahoney LT. Heart rate monitoring of physical activity in children and adolescents: the Muscatine Study. Pediatrics 1992;89(2):256–61.

26. Massey PS, Lieberman A, Batarseh G. Measure of activity level in mentally retarded children and adolescents. Am J Ment Defic 1971;76(2):259–61.

27. Janz KF. Validation of the CSA accelerometer for assessing children's physical activity. Med Sci Sports Exerc 1994;26:369–75.

28. Janz KF, Witt J, Mahoney LT. The stability of children's physical activity as measured by accelerometry and self-report. Med Sci Sports Exerc 1995;27(9):1326–32.

29. Freedson PS, Evenson S. Familial aggregation in physical activity [published erratum appears in Res Q Exerc Sport 1992; 63(4):453]. Res Q Exerc Sport 1991;62(4):384–9.

30. Sallis JF, Patterson TL, Morris JA, Nader PR, Buono MJ. Familial aggregation of aerobic power: the influence of age, physical activity, and body mass index. Res Q Exerc Sport 1989;60:318–24.

31. Sallis JF, Condon SA, Goggin KJ, Roby JJ, Kolody B, Alcaraz JE. The development of self-administered physical activity surveys for 4th grade students. Res Q Exerc Sport 1993;64:25–31.

32. Craig SB, Bandini LG, Lichtenstein AH, Schaefer EJ, Dietz WH. The impact of physical activity on lipids, lipoproteins, and blood pressure in preadolescent girls. Pediatrics 1996;98(3 Pt 1): 389–95.

33. Godin G, Shephard RJ. Normative beliefs of school children concerning regular exercise. J School Health 1984;54:443–5.

34. Aaron DJ, Kriska AM, Dearwater SR, Anderson RL, Olsen TL, Cauley JA, et al. The epidemiology of leisure physical activity in an adolescent population. Med Sci Sports Exerc 1993;25(7): 847–53.

35. Brener ND, Collins JL, Kann L, Warren CW, Williams BI. Reliability of the Youth Risk Behavior Survey questionnaire. Am J Epidemiol 1995;141(6):575–80.

36. Andersen LB, Haraldsdottir J. Tracking of cardiovascular disease risk factors including maximal oxygen uptake and physical activity from late teenage to adulthood: an 8-year follow-up study. J Intern Med 1993;234:309–15.

37. Bouchard C, Tremblay A, LeBlanc C, Lortie G, Savard R, Theriault G. A method to assess energy expenditure in children and adults. Am J Clin Nutr 1983;37:461–7.

38. Weston AT, Petosa R, Pate RR. Validation of an instrument for measurement of physical activity in youth. Med Sci Sports Exerc 1997;29:138–43.

39. Sallis JF, Buono MJ, Roby JJ, Micale FG, Nelson JA. Seven-day recall and other physical activity self-reports in children and adolescents. Med Sci Sports Exerc 1993;25:99–108.

40. Last JM, editor. A dictionary of epidemiology. New York: Oxford Univ. Press, 1983.

41. LaPorte RE, Montoye HJ, Caspersen CJ. Assessment of physical activity in epidemiologic research: problems and prospects. Public Health Rep 1985;100:131–46.

42. Schoeller DA, Ravussin E, Schutz Y, Acheson KJ, Baertschi P, Jejuier E. Energy expenditure by doubly labeled water: validation in humans and proposed calculation. Am J Physiol 1986; 250:R823–30.

43. Goran MI, Carpenter WH, Poehlman ET. Total energy expenditure in 4- to 6-yr-old children. Am J Physiol 1993;264(Endocrinol Metab 27):E706–11.

44. Hovell MF, Bursick JH, Sharkey R, McClure J. An evaluation of elementary students' voluntary physical activity during recess. Res Q 1978;49:460–74.

45. Corbin CB, Pletcher P. Diet and physical activity patterns of obese and non-obese elementary school children. Res Q 1968;39:922–8.

46. Klesges LM, Klesges RC. The assessment of children's physical activity: a comparison of methods. Med Sci Sports Exerc 1987; 19:511–7.

47. Mukeshi M, Gutin B, Anderson W, Zybert P, Basch C. Validation of the CALTRAC® movement sensor using direct observation in young children. Pediatr Exerc Sci 1990;2:249–54.

48. Noland M, Danner F, DeWalt K, McFadden M, Kotchen JM. The measurement of physical activity in young children. Res Q Exerc Sport 1990;61:146–53.

49. Klesges RC, Woolfrey J, Vollmer J. An evaluation of the reliability of time sampling versus continuous observation data collection. J Behav Ther Exp Psych 1985;16:303–7.

50. Welk GJ, Corbin CB. The validity of the TRITRAC®-R3D activity monitor for the assessment of physical activity in children. Res Q Exerc Sport 1995;66(3):202–9.

51. Bray MS, Wong WW, Morrow JR Jr, Butte NF, Pivarnik JM. CALTRAC versus calorimeter determination of 24-h energy expenditure in female children and adolescents. Med Sci Sports Exerc 1994;26:1524–30.

52. Ballor DL, Burke LM, Knudson DV, Olson JR, Montoye HJ. Comparison of three methods of estimating energy expenditure: CALTRAC, heart rate, and video analysis. Res Q Exerc Sport 1989; 60:362–8.

53. Livingstone MB, Coward WA, Prentice AM, Davies PSW, Strain

JJ, McKenna PG, et al. Daily energy expenditure in free-living children: comparison of heart-rate monitoring with the doubly labeled water ($^2H_2$ ^{18}O) method. Am J Clin Nutr 1992;56:343–52.

54. LaPorte RE, Cauley JA, Kinsey CM, Corbett W, Robertson R, Black-Sandler R, et al. The epidemiology of physical activity in children, college students, middle-aged men, menopausal females and monkeys. J Chronic Dis 1982:35:787–95.

55. Saris WH, Binkhorst RA. The use of pedometer and actometer in studying daily physical activity in man. Part II: validity of pedometer and actometer measuring the daily physical activity. Eur J Appl Physiol 1977;37:229–35.

56. Bush PJ, Iannotti RJ, Zuckerman AE, O'Brien RW, Smith SA. Relationships among black families' cardiovascular disease risk factors. Prev Med 1991;20(4):447–61.

57. Huttunen NP, Knip M, Paavilainen T. Physical activity and fitness in obese children. Int J Obesity 1986;10:519–25.

58. Murphy JK, Alpert BS, Christman JV, Willey ES. Physical fitness in children: a survey method based on parental report. Am J Public Health 1988;78:708–10.

59. Finegan JA, Niccols GA, Zacher JE, Hood JE. The Play Activity Questionnaire: a parent report measure of children's play preferences. Arch Sex Behav 1991;20(4):393–408.

60. Murphy JK, Alpert BS, Dupaul LM, Willey ES, Walker SS, Nanney GC. The validity of children's self-reports of physical activity: a preliminary study. J Hum Hypertens 1990;4:130–2.

61. Simons-Morton BG, O'Hara NM, Parcel GS, Huang IW. Baronowski T, Wilson B. Children's frequency of participation in moderate to vigorous physical activities. Res Q Exerc Sport 1990;61: 307–14.

62. Simons-Morton BG, Taylor WC, Huang IW. Validity of the Physical Activity Interview and Caltrac with preadolescent children. Res Q Exerc Sport 1994;65(1):84–8.

63. Sallis JF, McKenzie TL, Alcaraz JE. Habitual physical activity and health-related physical fitness in fourth-grade children. Am J Dis Child 1993;147:890–6.

64. Sallis JF, Strikmiller PK, Harsha DW, Feldman HA, Ehlinger S, Stone EJ, et al. Validation of interviewer- and self-administered physical activity checklists for fifth grade students. Med Sci Sports Exerc 1996;28(7):840–51.

65. Baranowski T, Dworkin RJ, Cieslik CJ, Hooks P, Clearman DR, Ray L, et al. Reliability and validity of self report and aerobic activity: Family Health Project. Res Q Exerc Sport 1984;55: 309–17.

66. Jenner DA, Vandongen R, Beilin LJ. Relationships between blood pressure and measures of dietary energy intake, physical fitness, and physical activity in Australian children aged 11-12 years. J Epidemiol Commun Health 1992;46(2):108–13.

67. Wallace JP, McKenzie TL, Nader PR. Observed vs. recalled exercise behavior: a validation of a seven day exercise recall for boys 11 to 13 years old. Res Q Exerc Sport 1985;56:161–5.

68. Baranowski T. Validity and reliability of self-report measures of physical activity: an information processing perspective. Res Q Exerc Sport 1988;59:314–27.

69. Liu K, Stamler J, Dyer A, McKeever J, McKeever P. Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol. J Chronic Dis 1978;31:399–418.

70. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: Wiley, 1981.

71. Bland JM, Altman DG. Statistical method for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.

72. Willett W, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. Am J Epidemiol 1986;124:17–27.

73. Sallis JF, Buono MJ, Freedson PS. Bias in estimating caloric expenditure from physical activity in children: implications for epidemiological studies. Sports Med 1991;11:203–9.

74. van den Berg-Emons RJG, Saris WHM, Westerterp KR, van Baak MA. Heart rate monitoring to assess energy expenditure in children with reduced physical activity. Med Sci Sport Exerc 1996;4:496–501.