Limits to the measurement of habitual physical activity by questionnaires

R J Shephard

Despite extensive use over 40 years, physical activity questionnaires still show limited reliability and validity. Measurements have value in indicating conditions where an increase in physical activity would be beneficial and in monitoring changes in population activity. However, attempts at detailed interpretation in terms of exercise dosage and the extent of resulting health benefits seem premature. Such usage may become possible through the development of standardised instruments that will record the low intensity activities typical of sedentary societies, and will ascribe consistent biological meaning to terms such as light, moderate, and heavy exercise.

Accurate measurement of habitual physical activity is fundamental to both the epidemiological study of relations between physical activity and health and the recommendation of an appropriate pattern of physical activity to maintain good health. If small numbers of subjects are to be studied, activity patterns can be determined in many different ways, including direct calorimetry, the ingestion of doubly labeled water, the use of motion sensors, accelerometers, heart rate recorders, or oxygen consumption meters, direct observation of movements by a trained observer, or assessments of food intake. However, epidemiological studies are often concerned with rare events, and the physical activity of large populations must then be categorised in order to draw significant conclusions. In the past, reported occupation has been successfully used to classify levels of physical activity, but mechanisation, automation, and the skills of the ergonomist have together reduced the energy cost of most jobs to a point where an occupation based categorisation of activity is no longer of great value. A questionnaires has thus become the only feasible method of assessing habitual physical activity in large populations.

This review of recent literature looks critically at various problems with questionnaire assessment of the type, intensity, frequency, and duration of physical activity, and the environment in which it is performed. It explores how each of these variables may best be measured, summarising information on the reliability and validity of different types of instrument. It stresses that interrelations among the scores from rival questionnaires, and relations between individual scores and other measures of health status are at best moderate. It also notes that attempts to interpret the data quantitatively can result in quite large absolute errors. Finally, some practical lessons are drawn for the epidemiologist and those prescribing physical activity.

Types of Physical Activity

After many years of confusion, a consensus has now been reached on the definitions of physical activity, exercise, sport, recreation, occupational activity, and household chores. Physical activity comprises all types of muscular activity that increase energy expenditure substantially. Exercise is a regular and structured subset of physical activity, performed deliberately and with a specific purpose such as preparation for athletic competition or the improvement of some aspect of health. Concepts of sport still differ between North America and some European countries. In North America, sport necessarily implies an activity that involves competition, whereas in Europe it may include recreational activities such as walking or hiking. Some forms of sport—for example, fishing and motor racing—do not involve a great deal of physical activity, and others, such as ice hockey and baseball, may become a job rather than a voluntary form of activity. Competition can be a source of motivation and self-esteem for those who are successful, but it also increases the risk of both cardiovascular and musculoskeletal injury.

Recreational activity varies widely in its intensity, and the participant may attach value to the environment in which it is pursued. The workplace was once a major source of weekly energy expenditure, but in developed countries it has become a progressively less important component for most people. Household and other chores are a significant but sometimes largely overlooked component of the total weekly energy expenditure, particularly in full time caregivers.

For some purposes, it is useful to distinguish the predominant type of activity, and indeed some surveys have distinguished occupational from leisure activity, or occupational, sport, and leisure activity. However, most aspects of health depend on the total amount of activity that is performed, and provided that the individual types of activity are all recognised and included in a global assessment, primary interest attaches to this overall score.

Patterns of Physical Activity

Surveys have commonly focused on the intensity, frequency, duration, and total amount of physical activity performed. The relative proportions of...
aerobic and resistance activity and the environmental context have attracted less attention.

Intensity
The intensity of physical activity may be expressed in absolute terms, as an absolute expenditure relative to body mass or resting metabolism, or as a value relative to peak performance. From the viewpoint of physical conditioning, it has long been asserted that the last is the most important characteristic. In the case of aerobic training, attempts have thus been made to express data as percentages of maximal oxygen intake, as fractions of the heart rate reserve, or most recently as fractions of the oxygen transport reserve. Likewise, resistance activity has commonly been expressed as a fraction of the one repetition maximum contraction force for a given muscle group. In support of the use of relative units, the physiological response to any given absolute intensity of effort appears to be greater in those who are unfit, or who have low initial levels of cardiorespiratory and muscular function because of aging. The influence of relative intensity seems logical and is well supported by experimental data. Nevertheless, one important remaining issue is the contribution to the observed relation of a reversion of data towards the true mean value as subjects who have been sorted in terms of their fitness are reassessed.

Despite strong empirical arguments for expressing data relative to peak ability, many authors still focus on the absolute intensity of effort. For instance, subjects are asked to describe a typical speed of walking, jogging, or cycling. Reference tables are then used to convert such information into an approximate estimate of energy expenditure (kJ/min), oxygen consumption (litres/min or ml/min per kg), or metabolic activity relative to resting conditions (METs).

The latter approach is likely to encourage compliance with sufficient weight bearing activity to enhance bone health. The impact of such activity on physical condition and health depends on the muscle groups that are exercised, the forces developed as fractions of maximal force for those muscle groups, the number of repetitions of contractions per set, the number of sets performed, and recovery intervals between sets.

Environmental context
Environment is a rarely noted aspect of physical activity. However, activity may be performed indoors or outdoors, in air or in water, under the relaxed conditions of a beautiful resort, or during a brief lunch break in a noisy gymnasium. The environmental conditions may be hot, with a high humidity and much radiant heating, or cold with a high wind chill factor. The ground surface may be smooth, rough, snow covered, or icy. Many of these factors alter the energy cost of a given activity, and some (such as heat stress) interact with training responses and changes in immune function.

Environment also influences the extent of ultraviolet exposure, the liability to heat stress or cold injury, the risk of musculoskeletal injury, and the impact of a given bout of physical activity on psychological health (particularly the stimulation received by a person in boring work, or the relaxation experienced by a person who is overstressed). Nevertheless, it remains arguable that, if the intensity, frequency, and duration of activity are established reliably, the environment has relatively little influence on many aspects of the aerobic response.

IMPLICATIONS FOR QUESTIONNAIRE DESIGN AND ANALYSIS

Basic issues
Questionnaires vary greatly in their detail, the period surveyed, and the extent of supervision of respondents. Some investigators have made only a simple global classification of subjects—for example, active versus inactive—or have asked only a very few simple questions, presenting data as a three to five category scale, an arbitrary summary index (exercise units), or a simple continuous variable (for example, MET.min of activity per week). Care must be taken in interpreting ordinal scales, because intercategory increments of energy expenditure or total activity may not be uniform. Other studies have used lengthy forms that require up to an hour to complete, often with assistance from a trained observer. Intercorrelations among the scores obtained from various complex questionnaires, and the correlations between such scores and assessments based on very few items, are often very low (0.14–0.41). Further, perhaps because
subjects become bored and/or confused by lengthy instruments, some of the highest coefficients of reliability and validity are seen for simple questionnaires. A comparison of the Baecke and Tecumseh questionnaires concluded that the former yielded superior results because it was simpler. Likewise, test/retest correlations were 0.81 for the very simple Godin questionnaire and 0.93 for a simple activity rating, and the Godin questionnaire also fared better than more complex instruments relative to a Caltrac motion assessment of concurrent validity.

Questionnaires may examine activities during the past one to seven days, through the last month to (in some instances) an entire lifetime. If the recording period is less than one week, care must be taken to include both weekday and weekend activities. Population sampling should also be dispersed to take account of seasonal variations in activity patterns. Questionnaire responses depend on the perception, encoding, storage, and retrieval of information about previous physical activity; answers depend on the subject's age and the context of questioning. Because of limitations in human memory, the reliability of information generally decreases with the length of the period surveyed, and it is best to keep the reporting interval relatively short (no longer than three months); however, in advanced age, long term memory may be better preserved than recent recollection of activity patterns. The accuracy of responses may be helped by asking questions about a specific time of day (for example, “what do you do after supper?”). Interval response options—for example, less than twice a week, two to three times a week, four to five times a week—may elicit a higher apparent frequency of physical activity than open ended questions.

Montoye et al. insisted that subjective responses to the detailed Tecumseh questionnaire required considerable interpretation. There was good correspondence between the activity ratings of three judges, but self reports were unsatisfactory unless combined with an interview. Another study found underestimation of activity by self report in military officers, but a better agreement between self reports and interviewers in ordinary working men. However, the media-
tion of an interviewer has sometimes had surprisingly little impact on the total amount of activity reported. In one study, self reports showed somewhat less activity than an interview assessment, but the one year activity scores reported by the subject and an interviewer correlated closely (r = 0.83). Interobserver (r = 0.42–0.99) and intraobserver (r = 0.56–0.96) inconsistencies were also noted by these investigators.

Questionnaire responses can be influenced considerably by cultural factors, in part because the content of reported activities differs from one country to another, and in part because the manner of answering questions shows a cultural bias. Difficulties are particularly likely if a questionnaire has been translated into another language. Respondents may also be influenced by the social desirability of reporting particular behaviours. In general, people tend to overreport physical activity and underestimate sedentary pursuits such as watching television. Sims and associates noted specifically that people who had been encouraged to exercise reported a greater volume of physical activity than could be confirmed by heart rate data.

In young children and those with mental impairments, attempts must be made to deduce physical activity patterns from the questioning of guardians, but such estimates have poor reliability and validity. In the elderly, other problems may arise from impairments of vision and hearing, and disturbances of cognition.

Types of activity

The number and type of activities reported can be augmented substantially by the use of either cue cards or leading questions on the part of an interviewer. A Swiss study suggested the value of listing 70 activities that together accounted for 95% of the weekly energy expenditure; if subjects indicated the number of days each activity was performed, and a typical duration for the activity, the resulting score was held to correlate well with estimates based on a heart rate monitor (r = 0.76). However, it remains uncertain whether such prompting recalls activities that have been performed very rarely, or whether they are important components of the total picture that the subject has inadvertently overlooked.

If interest is focused on specific activities, difficulty may arise because subjects obtain most of their weekly physical activity from items that are not listed on the cue cards. There have been suggestions that this approach may underestimate the volume of activity performed by full time caregivers and the elderly. A recent study based on doubly labelled water found that elderly women tended to overestimate their involvement in high scoring components of housework.

Intensity

Questionnaires often express the intensity of physical activity semantically, using a Likert-type scale. Unfortunately, perceptions of the intensity of any stimulus depend on the experience and the stoicism of the person concerned. Some people are particularly prone to report symptoms. Reporting may also be influenced by the perceived desirability of a given response. In the case of sport, additional information on the intensity of activity may be derived from the level of competition, the number of training sessions a week, and the time required to perform a standard task such as swimming four lengths of a 25 m pool. At work, where all intensities of effort are relatively low, distinction may be drawn between portions of the day spent sitting, standing, walking, and lifting or carrying.

There have been attempts to anchor semantic descriptions of exercise intensity in physiological terms, as with the original Borg scale, where each unit of perceived intensity was intended to correspond to a 10 beats/min increase in heart rate. Over a typical 15–30 minute bout of endurance exercise, the average person will perceive a task in the aerobic training zone as moderately hard (a Borg rating of 12–14 units), corresponding in a middle aged adult to a heart rate of 120–140 beats/min). Other potential anchors of intensity include “exercise sufficient to induce moderate sweating” or “causing sufficient breathlessness to limit conversation.” However, such descriptions at best distinguish light from vigorous effort. They are again somewhat vulnerable to differences in symptom reporting, and in the case of sweating are affected by environmental temperatures.

A further issue is the probable need to measure very low levels of physical activity. Unfortunately, many questionnaires suffer from floor effects. For example, one widely used seven day recall instrument does not take account of activities that are less intense than brisk walking, or that have a duration of less than ten minutes. The shape of the dose/response curve remains unclear, but some recent research suggests that, particularly in the frail elderly and those who are extremely sedentary, health advantages may accrue from very low levels of physical activity that are unlikely to induce breathlessness, sweating, or an increase in aerobic fitness.

Many questionnaires focus on the absolute rather than the relative intensity of individual physical activities. For instance, subjects are asked to specify a typical speed of walking, jogging, or cycling. Using a table of energy costs, many (but not all) reported activities can be converted into an approximate estimate of the rate of energy expenditure (kJ/min), an intensity of metabolic activity relative to resting conditions (METs), or an oxygen consumption (ml/min/per kg).
Nevertheless, there are substantial interindividual and intraindividual variations in the energy cost of various activities, depending on the subject's age, sex, body mass, skill, and level of fatigue. For example, the pace of walking differs considerably between those who are undertaking the activities of daily living and those who are performing deliberate exercise. Moreover, the costs of some activities are either unknown, or have changed since the data were first collected. Finally, some authors have translated data to MET values by assuming a standard value of 4.19 kJ/min (1 kcal/min) for basal metabolism; in fact, values vary with age, sex, and body surface from about 3.47 to 2.55 kJ/min per m².

Measures of absolute and relative intensity of effort may give widely differing estimates of the prevalence of adequate physical activity. Translation of absolute data to a relative intensity of activity is possible only if the subject's maximal oxygen intake is known. Questionnaires do not normally provide such information. There have been suggestions that subjects have a fairly clear perception of their physical fitness and that moderately accurate predictions of fitness can be made from age, body mass, skinfold thicknesses, and global assessments of habitual physical activity without engaging subjects in an exercise test. Nevertheless, the generality of such prediction equations is questionable, and the confidence limits are so broad that it is difficult even to categorise a person's fitness status.

Some analyses have apparently confounded the intensity of energy expenditure with the total quantity of energy expended a week. For example, one recent review described those with an energy expenditure of >20 METs/h/week as “highly active”. However, this expenditure could have been reached through 13 h/week of occupational activity at an intensity of only 1.5 METs. Increases in the total weekly energy expenditure that are achieved by moderate or low intensities of activity can be important for some aspects of metabolic health, although data summarised as METs or kJ/week cannot answer questions about the importance of a given absolute or relative intensity of physical activity to the prevention of ischaemic heart disease.

**Frequency**

The frequency of activity is usually reported as times per week or times per month. This may be a reasonable approach when making an overall assessment of habitual activity—for example, the number of sessions of sweat inducing activity of 20 minutes duration or longer. However, if such an assessment is applied repeatedly to a wide range of individual activities such as walking, running, cycling, and swimming, subjects are liable to overestimate the total hours of activity that they perform in a week.

The frequency of many activities varies substantially, even over an interval of a few months, and unfortunately respondents are liable to indicate their highest recent or their desired proportions of aerobic and resistance activity. However, if a large population is to be examined, true population averages can be approximated by dispersing questionnaire assessments over an entire calendar year. In view of recent research on the value of split exercise sessions for frail patients, it may be useful to keep track of divided sessions during a given day, particularly in those with low levels of fitness.

**Duration**

The duration of some types of activity tends to be overreported, making it necessary to adjust reported data substantially in order to limit the total length of a subject's day to 24 hours. In addition to problems of exaggeration, the indicated minutes of attendance at a sports club may include time devoted to changing, refreshment, and socialising. Some of the largest overestimates of exercise duration come from the school gymnasium, where the majority of a 30 or 40 minute physical education class may be spent in listening to instructions and awaiting a turn to use a particular item of equipment. Problems of underreporting can also arise through failure to take account of brief periods of activity encountered during some forms of everyday activity.

In young children, questionnaire assessments of physical activity are greatly complicated by their propensity for repeated brief bouts of vigorous physical activity.

**Timing and overall duration of activity**

Given that people do not maintain a consistent exercise behaviour throughout their lifetime, it may be important to ascertain when activity has been performed. In terms of the heart, current physical activity seems the most important determinant of both fitness and health. Little or no benefit is found from former athletic or leisure activity. But in terms of the prevention of osteoporosis, the critical factor may be the maximisation of bone mass during early adult life, and the key to prevention of some neoplasms may lie in adequate activity during adolescence, when cell division is at a maximum. Unfortunately, attempts to determine either the amount or the intensity of physical activity performed many years previously have only limited reliability and validity (see below).

**Amount of physical activity**

The impact of physical activity on certain metabolic variables such as obesity, the risk of diabetes mellitus, and hypercholesterolaemia seems to depend mainly on the total amount of energy expended, and many reports have summarised activity levels in such terms. However, increases in energy expenditure are necessarily the product of the net intensity and duration of activity, and commonly the two variables are confounded. A large energy expenditure is usually accumulated because a person chooses to exercise at a relatively high intensity of effort. At least one analysis has suggested that energy expenditures accumulated in non-vigorous physical activity do not influence longevity.

**Aerobic versus resistance activity**

Few questionnaires have addressed the issue of the relative proportions of aerobic and resistance activity. Information on resistance exercise is likely to be available if the subject has undertaken some type of circuit training, but the extent of such activity is very difficult to determine if the main type of activity is the performance of normal daily activities. Investigators are currently exploring potential questionnaires that can assess the extent of resistance activity.

**Environmental issues**

To my knowledge, none of the existing questionnaires of personal physical activity habits explore the type of environment in which an individual normally undertakes physical activity. The type of environment has particular importance in the contexts of motivation and the psychological benefits of physical activity.

**RELIABILITY, VALIDITY, AND SENSITIVITY OF MEASURING INSTRUMENTS**

A number of questionnaires were used quite widely before issues of their reliability and validity had been addressed. The current number of questionnaires suggests that many do not yield either reliable or valid information. Estimates of the prevalence of limited physical activity among women of child bearing age in the United States have ranged from 3.9% to 39.0%, using questions from three surveys conducted by the National Center for Health Statistics, Likewise, in the behavioural risk factor surveillance system, the prevalence of...
moderate activity as assessed by differing algorithms ranged from 20% to 38%. The proportion of the population of the United States who appear to meet current fitness guidelines varies from 32% to 59%, depending on the test instrument and the scoring protocol used; the proportion meeting health related guidelines varies even more widely, from 4% to 70%. Factors contributing to this wide variability include not only personal characteristics (age, sex, and socioeconomic status) but also the use of prompting cards and/or questions, and the number of items included in the estimate—for example, leisure activity versus leisure + occupation + household + transportation.

Reliability
The reliability of a questionnaire reflects its ability to yield the same result if it is applied on a second occasion. If the test is administered by an observer, variance due to errors of subject reporting and true changes in activity patterns between the two assessments are compounded by interobserver and/or intraobserver errors. Looking retrospectively at the total activity accumulated over 20 years, one report noted an interobserver reliability coefficient of about 0.90, and an intraobserver coefficient (on an older group of subjects than the interobserver study) of 0.70. Appropriate statistical methods must be used in assessing the reliability and validity of questionnaire responses. Several authors have used \( \chi \) statistics. Booth and associates evaluated the World Health Organisation Health Behaviour in Schoolchildren (WHO HBSC) questionnaire, showing a 70% agreement between two way classifications of 13 and 15 year old children (active/insufficiently active) after an interval of two weeks. Other statistically acceptable alternatives include calculations of the coefficient of variation in response (SD/mean) or the coefficient of repeatability (= 2SD). A small scale analysis of minutes per week of moderate or greater leisure and occupational activity found respective coefficients of repeatability of 29.3 and 54.6 minutes on totals of about 150 and 450 minutes.

More commonly, reliability has been evaluated in terms of test/retest intraclass correlation coefficients. This is a less desirable approach, in part because the magnitude of correlation is influenced by the extent of interindividual variance within the data set. If the samples evaluated are uniformly sedentary, low coefficients of reliability and validity are to be expected. Of potentially recorded characteristics (intensity, frequency, and duration), intensity seems to be the least reliably reported. Problems arise from differing individual perceptions of a given absolute or relative intensity of effort, and a lack of agreement on MET values corresponding to vigorous, moderate, and light activity. The reliability of responses also varies with the interval between tests. Two week test/retest observations on the simple questionnaire of Godin and Shephard found an intraclass correlation coefficient of 0.94 for reports of strenuous activity (estimated intensity 9 METs), falling to only 0.46 for moderate activity (5 METs) and 0.48 for light activity (3 METs).

Most authors have looked at indices of total activity. Reliability diminishes with the length of the recall period. Again, this has been assessed by test/retest correlations. Lamb and Brodie found a two week coefficient of 0.86, and the five week coefficient for the Minnesota leisure time physical activity questionnaire was 0.88. Studies on the college alumni questionnaire found \( r \) values of 0.72 at one month, falling to 0.3–0.4 over 8–12 months. Other authors have reported coefficients of a similar order: 0.58–0.67 for the community health activities model program for seniors (CHAMPS) physical activity questionnaire over a six month interval, 0.53 for adolescents over one year, and 0.59 for a two to three year recall in the coronary artery risk development in young adults (CARDIA) study.

Lack of reliability is due in part to seasonal and/or temporal variations in physical activity patterns, but shortcomings of human memory are also an important problem. Thus, questionnaire responses show a variation of 50% or more, even if one year activity patterns are reassessed after an interval of a few days. The problem is particularly acute if intensities of effort are low.

The reliability of absolute scores has received little attention, although there have been suggestions that, if a questionnaire is completed on several occasions, subjects become less precise in their responses, and intraindividual variations in reported activity diminish.

Validity
The reliability of a questionnaire reflects its ability to yield the same result if it is applied on a second occasion. If the test is administered by an observer, variance due to errors of subject reporting and true changes in activity patterns between the two assessments are compounded by interobserver and/or intraobserver errors. This is a less desirable approach, in part because the magnitude of correlation is influenced by the extent of interindividual variance within the data set. If the samples evaluated are uniformly sedentary, low coefficients of reliability and validity are to be expected. Of potentially recorded characteristics (intensity, frequency, and duration), intensity seems to be the least reliably reported. Problems arise from differing individual perceptions of a given absolute or relative intensity of effort, and a lack of agreement on MET values corresponding to vigorous, moderate, and light activity. The reliability of responses also varies with the interval between tests. Two week test/retest observations on the simple questionnaire of Godin and Shephard found an intraclass correlation coefficient of 0.94 for reports of strenuous activity (estimated intensity 9 METs), falling to only 0.46 for moderate activity (5 METs) and 0.48 for light activity (3 METs).

Most authors have looked at indices of total activity. Reliability diminishes with the length of the recall period. Again, this has been assessed by test/retest correlations. Lamb and Brodie found a two week coefficient of 0.86, and the five week coefficient for the Minnesota leisure time physical activity questionnaire was 0.88. Studies on the college alumni questionnaire found \( r \) values of 0.72 at one month, falling to 0.3–0.4 over 8–12 months. Other authors have reported coefficients of a similar order: 0.58–0.67 for the community health activities model program for seniors (CHAMPS) physical activity questionnaire over a six month interval, 0.53 for adolescents over one year, and 0.59 for a two to three year recall in the coronary artery risk development in young adults (CARDIA) study.

Lack of reliability is due in part to seasonal and/or temporal variations in physical activity patterns, but shortcomings of human memory are also an important problem. Thus, questionnaire responses show a variation of 50% or more, even if one year activity patterns are reassessed after an interval of a few days. The problem is particularly acute if intensities of effort are low.

The reliability of absolute scores has received little attention, although there have been suggestions that, if a questionnaire is completed on several occasions, subjects become less precise in their responses, and intraindividual variations in reported activity diminish.

Validity
Quite a number of investigators have limited their questionnaire evaluations to an examination of reliability, neglecting the more important issues of validity and sensitivity of response. Validity has a number of components (content, predictive, concurrent, and construct). Physical activity questionnaires should ideally be validated in terms of their criterion validity (a combination of predictive and concurrent validity, indicating the correspondence of scores to a more precise assessment of the characteristic of interest—for example, the total volume of physical activity performed). However, given the absence of any widely accepted criterion of physical activity, reliance has usually been placed on construct validation against other observations that are linked with physical activity. One analysis also looked at an expression of concurrent validity, comparing reported physical activity with the stage of change in exercise behaviour.

Measurements of energy expenditure using doubly labelled water are commonly accepted as the optimum in construct validation. The within subject variation for this technique (analytical plus biological variation) is about 8%. However, the necessary analyses are costly (as much as US$600 for a single measurement), and the data at best provide a two week average of energy expenditure. Further, the energy cost of many activities varies substantially between people, and in some disease states data interpretation may be complicated by an increase in basal metabolism. Tests against doubly labelled water have yielded correlation coefficients of 0.68 for the Baecke total activity index, 0.57 for the activity scale from the Five City Project questionnaire, 0.64 for the Tecumseh estimate of total energy expenditure, and 0.79 in men and 0.68 in women for the physical activity scale for the elderly. Some form of motion sensor such as a pedometer or accelerometer has provided a second, cheaper construct. Such devices tend to underestimate walking and overestimate jogging activity, also failing to detect arm movements, resistance exercise, and the performance of external work. Motion sensor scores may also show only weak relations to maximal oxygen intake. A Dutch questionnaire showed correlations of 0.78 and 0.73 with 24 hour activity recall and pedometer measurements respectively. Two thirds of subjects were assigned to the corresponding activity tertile by each of these methods. Other authors have found coefficients of around 0.70. Seven day scores for a Japanese pedometer showed correlations of 0.68–0.69 with questionnaire scores, the latter overestimating total energy intake by 4.5% in men, but not in women. Likewise, the correlation with the Minnesota leisure time physical activity questionnaire was 0.69 when a motion sensor was waist mounted, although it fell to 0.43 when the device was attached to the leg. Other investigators have found weaker correlations, possibly because most of their subjects were sedentary; one week activity scores had a correlation of 0.36 with Caltrac motion sensor values, the coefficient relating three day portable accelerometer scores to the physical activity scale for the elderly was 0.49, the Baecke and pre-EPIC questionnaires showed correlations of only 0.22 with Caltrac motion sensor...
scores in elderly women, and correlations between the Tecumseh community questionnaire and triaxial accelerometer scores were in the range 0.26–0.47. In patients in which the intensity of activity is limited by chronic disease, the correlation with accelerometer scores or doubly labelled water measurements may become non-significant—for example, \( r = 0.14 \) in chronic obstructive pulmonary disease and \( r = 0.057 \) in peripheral vascular disease.

Other approaches to validation have included comparisons with information obtained from heart rate records, exercise logs or diaries, 24 hour activity recalls, fitness scores, food consumption, and health outcomes. Significant correlations with aerobic fitness should be observed only for vigorous, sweat producing activity. In keeping with this expectation, correlation coefficients have been largest for the hardest forms of activity. Testing correlations with physical work capacity, one study found a correlation of 0.55 for “very hard” leisure time activity and 0.48 for “hard” activity, but much weaker correlations for less intense activity. A second report examined correlations between exercise test scores and responses to the Minnesota leisure time physical activity questionnaire in Spanish women; correlations fell from 0.51 for heavy activity to 0.13 for moderate and 0.02 for light activity.

Harada and associates found correlations of 0.44–0.68 between scores on three types of activity questionnaire (CHAMPS questionnaire, physical activity survey for the elderly, and the Yale physical activity survey) and performance based tests of lower body functioning and endurance. In a relatively active population, an overall index derived from the Harvard alumni questionnaire showed a correlation of 0.52 with aerobic power, although two of the three elements in this index (stepping and walking scores) showed almost zero correlation with aerobic performance (respective \( r \) values of 0.02 and 0.01). The correlation between scores on the physical activity scale for the elderly and peak oxygen intake was only 0.20. Likewise, an interviewer administered seven day activity recall showed very low correlations with estimated maximal oxygen intake (0.34), resting heart rate (–0.09), and body mass index (–0.23). Correlations with treadmill measurements of maximal oxygen intake were 0.31 for an index based on running, walking, and jogging, 0.35 for a question based on frequency of sweating, and 0.29 for Paffenbarger’s leisure time activity index. Other correlations were for treadmill run times (0.41), for physical work capacity at a heart rate of 150 beats/min (0.08 and 0.10), for 1.6 km run times (–0.37), for submaximal treadmill scores (0.13), and for total energy intake (–0.10, 0.27, and 0.02). Nevertheless, it was possible to account for some 75% of the variance in accelerometer scores of patients with end stage renal disease using a combination of the physical activity scale for the elderly and the human activity profile.

Godin and Shephard developed discriminant functions to predict two way classifications of aerobic fitness and body fat content from responses to a simple activity questionnaire. Respective \( \kappa \) values were 0.30 and 0.17, with 69% and 66% of subjects being classified correctly.

Often, investigators have accepted validity coefficients of 0.3–0.5 relative to other direct or indirect measures of physical activity and energy expenditure. Thus Bailey-Mertz and associates concluded that the Duke activity status index was “a reasonable correlate of functional capacity,” given a coefficient of 0.31—that is, it described 9.6% of the total variance in functional capacity. A two year trial on nurses found “reasonably valid” measures: a test/retest correlation of 0.59, and a correlation of 0.60 between diary and questionnaire data. A third report suggested that a questionnaire gave a “reasonable” estimate of physical activity over the past year, even though scores showed no significant relations to physical fitness or body mass index.

Even in studies in which correlation coefficients have been relatively high, absolute estimates of physical activity have shown large errors. Thus, one comparison between the college alumnus questionnaire and pedometer scores gave respective estimates of daily walking distance as 2.3 (1.6) and 6.7 (2.6) km. The gross energy cost of stair stepping was also only 50% of values assumed in the questionnaire analyses. Nevertheless, the total physical activity (MET.min/week) indicated by the questionnaire was only a third of values found in a 48 hour physical activity log. Attempts to recall activity that had previously been assessed 11 years ago also led to a 41% increase in the estimated total weekly energy expenditures.

### Sensitivity

An effective questionnaire must be not only reliable and valid, but also sufficiently sensitive to detect relevant activity related differences in health status and programme related changes in patterns of habitual physical activity. In some studies, in which little effect of a programme has been seen, it is difficult to be certain whether the problem lies with the programme or the measure of habitual physical activity.

In general, questionnaires seem to be less sensitive than more objective instruments such as accelerometers. Thus, accelerometer scores suggested that patients with peripheral arterial disease had only 46% of the energy expenditure of controls (\( p < 0.001 \)), whereas applications of the health interview survey and the Stanford seven day activity recall to the same subjects suggested values that were 73% (\( p = 0.128 \)) and 93% (\( p = 0.454 \)) of controls respectively. In Britain, civil servants classed as active had a significantly greater daily food intake than those classed as inactive. Likewise, there were 12–23% differences of shuttle run score between “active” and “inadequately active” adolescents, using the WHO HBSC questionnaire.

Blair and associates tested the ability of a seven day activity recall to detect associations between changes in energy expenditure and gains of fitness over a 12 month trial. They found correlations of 0.33 for maximal oxygen intake, –0.50 for body fatness, and 0.32 for high density lipoprotein cholesterol.

Small to moderate size effects (0.38–0.64) were noted when one questionnaire was used to evaluate the impact of a six month programme promoting physical activity. A much larger effect size of 1.68 was shown when pedometers were used to evaluate a four week programme for patients with type 2 diabetes mellitus.

### CONCLUSIONS AND IMPLICATIONS FOR EPIDEMIOLOGY AND EXERCISE PRESCRIPTION

What practical conclusions can the epidemiologist and those formulating exercise prescriptions draw from this review? Irrespective of the questionnaire chosen, the data probably have limited reliability and validity relative to a laboratory measure of physical activity. If, as is commonly the case, the need is to calculate risk ratios for two or three different volumes of habitual physical activity—for example, high, moderate, and low—the use of a large number of subjects commonly reduces problems resulting from imprecise classification and allows demonstration of activity related benefits of reduced morbidity and mortality and enhanced health. Nevertheless, misclassification reduces the apparent magnitude of any benefits from physical activity. This is an important reason why a three level classification of aerobic fitness (assessed accurately in the laboratory) apparently has a larger influence on all cause mortality than a three level classification of habitual physical activity.

If large populations are examined, categorisation may also show significant dose/response relations. However, the absolute energy expenditures corresponding to light, moderate, and vigorous effort remain unclear, and attempts to interpret questionnaire data in a quantitative sense are generally...
Take home message

Despite extensive use over 40 years, physical activity questionnaires still show limited reliability and validity. Measurements have value in indicating conditions where an increase in physical activity would be beneficial, and in monitoring changes in population activity. However, attempts at detailed interpretation in terms of exercise dosage and the extent of resulting health benefits seem premature. Such usage may become possible through the development of standardised instruments that will record the low intensity activities typical of sedentary societies, and will ascribe consistent biological meaning to terms such as light, moderate and heavy exercise.

unwarranted. Plainly, there is need for international consensus on the wording of questionnaires and the methods of analysis and interpretation. Both data interpretation and comparisons between studies would be greatly facilitated if individual observers used a reference standard such as doubly labelled water or even heart rate recordings on a small sample of their subjects, to clarify the average energy expenditures characteristic of their activity categories.

The choice of questionnaire depends ultimately on the purpose of the investigator, and the available resources of time, funding, and skilled personnel. Nevertheless, for many purposes an accurate but simple classification of activity levels may be more appropriate than an attempt at estimating overall energy expenditures. Care must be taken to avoid bias when distinguishing between categories of activity intensity or volume. Cut point bias may be introduced because categorisation is adjusted to fit sample distribution—for example, the use of tertiles— or to maximise statistical significance. In particular difficulty is experienced in detecting and assessing low levels of physical activity, and given that this is the most prevalent form of activity in the general North American population, attention must be focused on developing better methods to assess low intensity effort.

Measurement errors assume particular significance in discussions of dose/response relations and recommendations of a minimum dose of physical activity to optimise population health. Many authors claim gains from quite moderate intensities of effort (as little as 40–50% of oxygen consumption or heart rate reserve) of aerobic exercise. Others have argued that, after adjustment for total energy expenditure and other confounding variables, no significant benefit is obtained unless the intensity of effort exceeds an absolute level of 6 METs and the total energy expenditure exceeds 2 MJ/week. Nevertheless, proponents of high intensity exercise found a trend (p<0.07) to benefit from moderate activity (intensity >4 METs), and it seems likely that if their questionnaire classification of effort intensity had been more precise, this trend would have been significant.

Likewise, much, if not all, of the apparent energy expenditure threshold of 2.1–4.2 kJ/week would disappear if account were taken of systematic errors in the estimation of energy costs and data were expressed as net rather than gross energy expenditures.

Despite the problems outlined in this review, physical activity questionnaires have practical value in indicating conditions where an increase in physical activity would be beneficial and in monitoring changes in population activity. Attempts at more detailed interpretation in terms of exercise dosage and the extent of resulting health benefits seem premature. However, such usage may become possible through the development of standardised instruments that will record the low intensity activities typical of sedentary societies, and will ascribe consistent biological meaning to terms such as light, moderate, and heavy exercise.

REFERENCES

Measurement of physical activity


Numerous studies on physical activity and health have been published. There are so many tools used to measure physical activity that the problem of validity, reliability, and sensitivity arises. Despite the limitations of questionnaires, they are often used, especially for large populations or studies investigating influence of lifetime physical activity. This paper gives an overview of the limitations and points out the difficulties associated with measurement of physical activity. Different types and patterns of physical activity are detailed, which are used to provide various indicators. It would also be interesting to have an inventory of these indicators and their use depending on the aims of the study.

A Vuillemin, Université Henri Poincaré - Nancy 1, Faculté du Sport, 30 rue du Jardin Botanique, 54600 Villers-les-Nancy, France; Anne.Vuillemin@staps.uhp-nancy.fr

If you wish to comment on any article published in the British Journal of Sports Medicine you can send an eLetter using the eLetters link at the beginning of each article. Your response will be posted on British Journal of Sports Medicine online within a few days of receipt (subject to editorial screening).

www.bjsportmed.com