

Methodologic issues in measuring physical activity and physical fitness when evaluating the role of dietary supplements for physically active people¹⁻³

William L Haskell and Michaela Kiernan

ABSTRACT Physical activity and physical fitness are complex entities comprising numerous diverse components that present a challenge in terms of accurate, reliable measurement. Physical activity can be classified by its mechanical (static or dynamic) or metabolic (aerobic or anaerobic) characteristics and its intensity (absolute or relative to the person's capacity). Habitual physical activity can be assessed by using a variety of questionnaires, diaries, or logs and by monitoring body movement or physiologic responses. Selection of a measurement method depends on the purpose of the evaluation, the nature of the study population, and the resources available. The various components of physical fitness can be assessed accurately in the laboratory and, in many cases, in the field by using a composite of performance tests. Most coaches and high-level athletes would accept as very beneficial a dietary supplement that would increase performance in a competitive event by even 3%; for example, lowering a runner's time of 3 min, 43 s in the 1500 m by 6.7 s. To establish that such small changes are caused by the dietary supplement requires carefully conducted research that involves randomized, placebo-controlled, double-blind studies designed to maximize statistical power. Statistical power can be increased by enlarging sample size, selecting tests with high reliability, selecting a potent but safe supplement, and maximizing adherence. Failure to design studies with adequate statistical power will produce results that are unreliable and will increase the likelihood that a true effect will be missed. *Am J Clin Nutr* 2000;72(suppl):541S-50S.

KEY WORDS Physical activity, exercise, exercise training, physical fitness

INTRODUCTION

Dietary supplements can be used by physically active people to increase their physical performance (physical fitness), improve their health, or reduce the potentially negative consequences of physical activity (eg, injury, chronic fatigue, or suppressed immune function). To appropriately assess these effects, reliable and accurate measures of physical activity, physical fitness, and health-related outcomes must be made. All of these outcomes are complex entities consisting of several different characteristics or components that must be considered individually, depending on the specific scientific or clinical questions being addressed. Because of the numerous unsubstantiated claims about the

performance-enhancing effects of various dietary supplements, research in this area must be performed to the standards required for funding by the National Institutes of Health or for product approval by the US Food and Drug Administration.

Presented in this article are some of the key issues that need to be considered in measuring physical activity and physical fitness in physically active people who are using dietary supplements. Included are definitions of some key terms, a brief overview of the measurement of physical activity and physical fitness, and some issues in research design that are related to the measurement of the effects of dietary supplements on physical performance.

DEFINITIONS

The terms *physical activity*, *physical fitness*, and *health* have been assigned numerous definitions over the past few decades. Agreement on what the terms mean is important when considering their relation to or interaction with the use of dietary supplements. For our purposes, the terms are defined in relation to 1) the enhancement of physical performance or health as a result of changes in physical activity or exercise and 2) the use of dietary supplements.

Physical activity

Physical activity is defined as any bodily movement produced by the contraction of skeletal muscle. This relatively well understood biomechanical or biochemical process leads to a complex set of responses in the body that have a variety of health- and performance-related dimensions, the relation of which varies depending on the characteristics of the activity and the specific health outcome. Physical activity can be categorized by several variables, including type and intensity.

Muscle contraction has both mechanical and metabolic properties and thus can be classified by either of those categories,

¹From the Stanford Center for Research in Disease Prevention, Stanford University, Palo Alto, CA.

²Presented at the workshop Role of Dietary Supplements for Physically Active People, held in Bethesda, MD, June 3-4, 1996.

³Address reprint requests to WL Haskell, Stanford Center for Research in Disease Prevention, Stanford University, 730 Welch Road, Suite B, Palo Alto, CA 94304. E-mail: bhaskell@scrdp.stanford.edu.

TABLE 1
Glossary of terms

Term	Definition
Agility	A skill-related component of physical fitness that relates to the ability to rapidly change the position of the entire body in space with speed and accuracy ¹
Balance	A skill-related component of physical fitness that relates to the maintenance of equilibrium while stationary or moving ¹
Body composition	A health-related component of physical fitness that relates to the relative amounts of muscle, fat, bone, and other vital parts of the body ¹
Cardiorespiratory endurance	A health-related component of physical fitness that relates to the ability of the circulatory and respiratory systems to supply fuel during sustained physical activity and to eliminate fatigue products after supplying fuel ¹
Coordination	A skill-related component of physical fitness that relates to the ability to use the senses, such as sight and hearing, together with body parts in performing motor tasks smoothly and accurately ¹
Flexibility	A health-related component of physical fitness that relates to the range of motion available at a joint ¹
Muscular endurance	A health-related component of physical fitness that relates to the ability of muscle groups to exert external force for many repetitions or successive exertions
Muscular strength	A health-related component of physical fitness that relates to the amount of external force that a muscle can exert ¹
Physical activity	Any bodily movement produced by the contraction of skeletal muscle
Physical fitness	A set of attributes that people have or achieve that relates to the ability to perform physical activity
Power	A skill-related component of physical fitness that relates to the rate at which one can perform work
Reaction time	A skill-related component of physical fitness that relates to the time elapsed between stimulation and the beginning of the reaction to it ¹
Speed	A skill-related component of physical fitness that relates to the ability to perform a movement within a short period of time ¹

¹ Adapted from Corbin et al (4).

a situation that has caused some confusion. Typically, mechanical classification stresses whether the muscle contraction produces movement of the limb: isometric (same length) or static exercise if there is no movement, and isotonic (same tension) or dynamic exercise if there is movement. In addition, muscle contraction can be either concentric (shortening of the muscle fiber) or eccentric (lengthening of the muscle fiber). The metabolic classification involves the availability of oxygen for the contraction process and includes aerobic (oxygen available) or anaerobic (without oxygen) processes. Whether an activity is aerobic or anaerobic depends primarily on its intensity. Most activities involve both static and dynamic contractions as well as aerobic and anaerobic metabolism. Thus, activities tend to be classified by their dominant features.

The intensity of an activity can be described in both absolute and relative terms. In absolute terms, intensity is either the mag-

nitude of the increase in energy required to perform the activity or the force produced by the muscle contraction. The increase in energy is usually determined by measuring the increase in oxygen uptake, which is expressed in units of oxygen or converted to a measure of heat or energy expenditure (kJ). The force of the muscle contraction is measured by how much weight is being moved or the force exerted against an immovable object and is expressed in kg or lb. In relative terms, the intensity of the activity is expressed in relation to the capacity of the person performing the activity. For energy expenditure, the intensity is usually expressed as a percentage of the person's aerobic capacity (percentage of maximal oxygen uptake, or $\dot{V}O_{2max}$). Because there is a linear relation between the increase in heart rate and the increase in oxygen uptake during dynamic exercise, the percentage of maximal heart rate or heart rate reserve (maximal heart rate minus resting heart rate) is also used as an expression of exercise intensity relative to the person's capacity. For muscle force, the relative intensity of the contraction is expressed as a percentage of the maximal force that can be generated for that activity (percentage of maximal voluntary contraction or percentage of one-repetition maximum).

Exercise (or exercise training)

Exercise and physical activity have generally been used interchangeably to represent movement produced by the contraction of skeletal muscle. It is more precise to say, however, that exercise (or exercise training) is a subcategory of physical activity or "physical activity that is planned, structured, repetitive, and purposive in the sense that improvement or maintenance of one or more components of physical fitness is the objective" (1, page 126). One problem with this definition is that many activities may be classified as both exercise and not exercise. For example, for one person, a brisk walk to work has the sole objective of transportation and thus is not exercise, but for another person the same walk may have a goal of reducing adiposity and thus be classified as exercise. It is recommended that the term *exercise training* be used when activity is performed for the sole purpose of enhancing physical fitness.

Physical fitness

Physical fitness has been defined in various ways (2). For purposes here, it is "a set of attributes that people have or acquire that relates to their ability to perform physical activity" (3, page 3). Being physically fit has been defined as "the ability to carry out daily tasks with vigor and alertness, without undue fatigue and with ample energy to enjoy leisure-time pursuits and to meet unforeseen emergencies" (3, page 5). This definition aptly describes what should be achieved from a program that promotes physical fitness. Although characteristics such as vigor, fatigue, alertness, and enjoyment are not easily measured, other measurable components of fitness can be used to assess a person's health or performance status on several different attributes (**Table 1**).

To define more accurately the outcomes of physical fitness programs for improving health rather than maintaining or enhancing physical or athletic performance, the concept of performance-related fitness compared with health-related fitness evolved (5). However, although a clear separation between the health- and performance-related components of physical fitness has been proposed (1), such a separation is not always possible. For example, cardiorespiratory endurance and muscle strength are highly important components of both kinds of fitness. In



TABLE 2
Components of physical fitness and their relation to physical performance and health¹

Contribution to health			Components of fitness	Contribution to performance		
High	Medium	Low		Low	Medium	High
-----			Cardiorespiratory endurance	-----		
-----			Skeletal muscle endurance	-----		
-----			Skeletal muscle strength	-----		
-----			Speed	-----		
-----			Flexibility	-----		
-----			Agility	-----		
-----			Balance	-----		
-----			Reaction time	-----		
-----			Body composition	-----		

¹The magnitude of the contribution will vary depending on the specific sport or activity being performed or the specific measure of health being considered.

Table 2, the contribution of each of the components of physical fitness to health- and performance-related fitness are qualitatively rated. As shown, most components contribute to both performance and health status. The magnitude of the contribution of any one component depends on the specific objective. For a gymnast, balance, agility, and power are extremely important, whereas cardiorespiratory endurance, skeletal muscle endurance, and body composition are vital for a distance runner. Moreover, an increase in muscle strength has little health benefit for healthy young women, but may be critical for a frail elderly woman who is at risk of falling and suffering an osteoporotic fracture.

Health

The 1988 International Consensus Conference on Physical Activity, Physical Fitness, and Health (6) defined health as "...a human condition with physical, social, and psychological dimensions, each characterized on a continuum with positive and negative poles. Positive health is associated with a capacity to enjoy life and to withstand challenges; it is not merely the absence of disease. Negative health is associated with morbidity and, in the extreme, with premature mortality" (page 84). Thus, when considering the role of physical activity or dietary supplements in promoting health, one needs to consider psychological well-being as well as physical health and reject the notion that simply being free of disease is optimal health.

MEASUREMENT OF PHYSICAL ACTIVITY AND PHYSICAL FITNESS

The measurement of physical activity and physical fitness in studies designed to determine their relation to health status and performance developed throughout the 20th century (2). Major reviews covering the issues involved in obtaining accurate and reliable measurements have been published (7)(9). When designing studies to evaluate the effects of dietary supplements on physical performance and health or the interaction of supplements with exercise training, it is important to understand the strengths and weakness of each of the various methods (10).

Physical activity

Physical activity is a complex and not easily measured set of behaviors. Numerous approaches have been used to assess physical activity or change in activity in studies in which health status or performance is the primary outcome. Self-reported surveys are used most frequently; other approaches have included job classification, behavioral observation, motion sensors, physio-

logic markers (eg, heart rate, doubly labeled water), and indirect and direct calorimetry. Most data that support a relation between physical activity status and clinical health outcomes were collected by using job classification (11) or a self-reported survey (12, 13). The other approaches noted above were typically used in smaller observational or intervention studies or to validate self-reported surveys (14, 15).

Self-reported surveys

To determine the relation between physical activity and health, researchers must use instruments that reliably assess habitual physical activity in the target population. Most of the scientifically sound data relating physical activity to morbidity and mortality were derived from prospective observational studies that used self-reported surveys such as diaries, logs, recall questionnaires, global self-reports, and quantitative histories (8, 16)(18). Surveys are frequently used because they are practical for assessing physical activity in large populations and have relatively low study and respondent costs (1, 19, 20).

Diaries

Diaries generally provide a detailed accounting of virtually all physical activity performed, normally within a single day. The summary index from a diary is typically a kJ score derived by summing products obtained by multiplying time spent in a given activity by an estimated rate of energy expenditure for that activity. When diary scores were compared with either indirect calorimetry or energy intake, they were shown to be accurate indexes of daily energy expenditure. Unfortunately, diaries tend to be used for time frames of 1–3 d, raising questions about how well they represent an individual’s long-term physical activity pattern (19). In addition, diaries require intensive effort by subjects and may even influence them to change their physical activities while being monitored (16, 19). In addition, diaries produce vast amounts of data, especially when multiple days are monitored, thereby requiring additional costs for data processing.

Physical activity logs

Activity logs provide an ongoing record of a subject’s participation in certain types of physical activity (21). The time of onset and cessation of physical activity may be recorded either immediately after or shortly after participation. In other instances, the recording is more conveniently recalled and recorded at the end of the day. Logs differ from diaries in that each behavior during the day is generally not recorded. Logs can demand too much time and be inconvenient for subjects to



complete accurately, and they can influence the subject's behavior. On the other hand, they can be very useful for recording specific activities such as participation in an exercise training program.

Recall surveys

Recall surveys are less likely to influence physical activity behavior and generally require less effort by the respondent than do either diaries or logs. Remembering details of prior participation in physical activity, which may be substantial, requires the greatest effort, especially among older persons or patients with cognitive deficits (22). Recall surveys have been used for time frames of 1 wk, 1 mo, 1 y, and even for lifetime physical activity (23, 24). Either precise details about physical activity or more general estimates of usual or typical participation in physical activity can be ascertained for the time frame of interest.

Retrospective quantitative history

This is the most comprehensive form of physical activity survey and generally requires specific detail for time frames of up to 1 y (19). If the time frame is long enough, the quantitative history can adequately represent seasonal physical activity. Both the Minnesota Leisure-Time Physical Activity Questionnaire and the Tecumseh Questionnaire obtain information on the average frequency and duration of participation over the prior year by using a specific list of physical activities (25, 26). Unfortunately, obtaining the data collected by the quantitative history places a large burden on respondents to remember all the details and also generates expenses for administering the survey, training the interviewers, ensuring quality control, and processing data (19).

Global self-report

The global self-report provides a self-assessment of an individual's physical activity relative to other persons in general or to those of a similar age and sex. This approach was used in the National Health Interview Survey 2 decades ago (27). The global self-report is easy to use and tends to represent participation in vigorous physical activity (10, 28, 29). However, when groups that vary by age or sex are compared, very different physical activity profiles may be observed among persons reporting the same self-assessed rating (28).

Various researchers have emphasized the development of self-report surveys for older persons (30–32) and adolescents or children (33, 34). These questionnaires provide reliable and valid methods for classifying elderly persons into physical activity groups (low to high), but they have not been shown to accurately or reliably measure changes in physical activity, especially low- to moderate-intensity activity. Obtaining highly accurate and reliable self-reported measures of physical activity in children has proven to be difficult because of their poor recall of activity intensity and duration (33).

Motion sensors and physiologic monitoring

Directly measuring physical activity by physiologic monitoring or motion sensors offers a potential advantage over self-reported data by reducing bias from poor memory and overreporting or underreporting. Limitations include the cost of high-quality monitors and the burden placed on subject and staff. Both the monitoring of physiologic processes related to physical activity, particularly heart rate, and mechanical or electronic sensors (pedometers, movement counters, and accelerometers) have been used in small-scale studies but not in large observational trials

with clinical events as outcomes. In addition, these monitors have been used to validate various self-reported surveys.

Heart rate. Monitoring heart rate can provide a continuous recording of a physiologic process that potentially reflects both the duration and the intensity of physical activity. Heart rate is typically used to estimate physical activity as energy expenditure (oxygen uptake), based on the assumption of a linear association between heart rate and energy expenditure. Heart rate measured during daily activities is thus used to establish energy expenditure. One major disadvantage of heart rate monitoring is the need to calibrate each individual; another limitation is that during low-intensity exercise the relation between exercise intensity and heart rate is frequently not linear.

Other approaches to the use of heart rate as a measure of physical activity have been suggested. For example, researchers have used the percentage of time spent during daily activities in various ranges of heart rate (35), the difference between mean daily heart rate and resting heart rate (36), and the integration of the area under the curve of heart rate versus time adjusted for resting heart rate (37). Heart rate alone may not be a suitable surrogate for determining level of physical activity, in that other factors such as psychological stress or changes in body temperature can significantly influence heart rate throughout the day.

Motion sensors. Pedometers, the original motion sensor for measuring physical activity, were designed to count steps and thus provide a potentially useful measure of distance walked or run. However, the high variability among pedometers and the lack of a stable calibration mechanism make them unsuitable for estimating physical activity in either laboratory or field research (38, 39). Electronic motion sensors have overcome much of the lack of standardization and poor quality control associated with mechanical pedometers. Devices used by various investigators include the Large-Scale Integrated Activity Monitor (40), the Caltrac Personal Activity Computer (Caltrac) (41), and the Vitalog monitor (42). The output from these monitors has been significantly correlated with energy expenditure assessed by indirect calorimetry during walking and running on the treadmill (43, 44), stationary cycling, walking over a measured course, and simulated activities of daily living (eg, lifting and carrying objects, sweeping) (36, 37, 39, 45, 46). Direct validation of the Caltrac shows low to moderate associations with physical activity records completed over the course of 1 y (47). Simultaneously recording the heart rate and motion from sensors on several parts of the body and calibrating each individual's heart rate and motion sensor output versus oxygen uptake for various activities can provide an accurate estimate of the energy expenditure profile from physical activity (44, 48). More advanced hardware and software are needed to make such approaches useful for studies measuring health outcomes of physical activity.

Doubly labeled water. By using 2 stable isotopes ($^2\text{H}_2\text{O}$ and H_2^{18}O), researchers can calculate the rate of carbon dioxide production in humans over days or weeks. Subjects drink a specified amount of these isotopes according to their body weight, after which their loss from the body is tracked by analysis (using a mass spectrometer) of isotopes in urine samples every few days. From these data, oxygen uptake and energy expenditure can be calculated. This technique has the advantage of obtaining objective data with little effort by subjects; its disadvantages include a relatively high cost and the inability to determine the type, intensity, frequency, or duration of any single bout of activity. This



technique has been shown to be accurate when compared with indirect calorimetry (49, 50).

Physical fitness

Measurements of the various health-related components of physical fitness have been developed and, in some cases, standardized, with good to excellent accuracy and reliability.

Cardiorespiratory endurance

In studies investigating the primary or secondary prevention of cardiovascular diseases, the major component of physical fitness that has been related to cardiovascular health or risk has been cardiorespiratory fitness or capacity (also referred to as cardiovascular, aerobic, or endurance fitness or capacity). Although other components of physical fitness, such as muscle strength or endurance, may relate to some aspects of cardiovascular health, few data document these relations.

One of the major reasons for measuring cardiovascular fitness in studies of the relation between physical activity and health is that habitual physical activity status is one of the major determinants of cardiovascular fitness. Other determinants include age, sex, heredity, medical status, and selected health-related behaviors (51). Thus, tests of cardiovascular fitness can be used as objective, surrogate measures of physical activity status with the understanding that factors other than activity influence the results. The magnitude of the effects of these other factors is generally reduced when changes in fitness are measured to verify changes in activity status.

The gold standard, or criterion measure, of cardiorespiratory fitness is maximal oxygen uptake or power ($\dot{V}O_2$ max). Measured in healthy persons during large-muscle, dynamic activity such as walking, running, or cycling, it is primarily limited by the oxygen transport capacity of the cardiovascular system (52). The most accurate assessment of $\dot{V}O_2$ max is made by measuring expired air composition and respiratory volume during maximal exertion. This procedure requires relatively expensive equipment, highly trained technicians, and time and cooperation from the subject, all of which make the procedure difficult for large-scale studies. This approach has primarily been limited to small-scale ($n < 200$) training studies but was used to assess cardiovascular fitness in a community-based sample of men and women in the Tecumseh Study (25) and with 396 men and women in the Stanford-Sunnyvale Health Improvement Project (21).

Because in the adult population the interindividual variation in mechanical and metabolic efficiency is quite low for standard testing activities, such as walking or running on a motor-driven treadmill or cycling on a stationary ergometer, oxygen uptake can be accurately estimated from the rate of work (speed, grade, and resistance) (53). Thus, $\dot{V}O_2$ max can be estimated from the peak exercise intensity during a maximal exercise test. This procedure requires an accurately calibrated exercise device, careful adherence to a specific protocol, and good cooperation by the subject. It has been used in numerous exercise training studies for evaluating the effects of exercise on cardiovascular risk factors and performance, in secondary prevention trials after hospitalization for myocardial infarction, and in a few large-scale observational studies, such as those conducted by the Institute for Aerobic Research (54) and the CARDIA project (55).

Having a subject perform any maximal test to assess cardiorespiratory fitness carries a substantial burden for both the subject and examiner. For the subject, the burden includes time, effort, and risk. To reduce this burden, various submaximal exer-

cise testing protocols have been developed and used in numerous observational and intervention studies for evaluating the relations between physical activity, cardiovascular fitness, and cardiovascular health. In most protocols, the estimate of cardiovascular fitness is made from the response of the heart rate to a set work rate or workloads, and data from the submaximal response are used to extrapolate to a predicted $\dot{V}O_2$ max. The underlying assumptions in this procedure are that a linear relation exists between heart rate and oxygen uptake and that the subject's maximal heart rate can be estimated reasonably accurately. Both assumptions are adequately met when a large sample of healthy adults is tested by using a standardized protocol. In some cases, no extrapolation to maximal values is performed, and an individual's cardiovascular fitness is expressed as the heart rate at a set workload (eg, heart rate at 5 km/h or at 100 W) or the workload required to reach a specific submaximal heart rate (eg, workload at a heart rate of 120 beats/min). These submaximal tests have been performed using motor-driven treadmills, cycle ergometers, and steps.

Another approach for assessing cardiorespiratory fitness has been field testing, where the performance of subjects who usually walk, jog, or run a specified time or distance is converted to an estimate of $\dot{V}O_2$ max or aerobic power (56). These procedures have frequently been used for children, for young adults, or for groups that have occupation-related physical fitness requirements, such as military and emergency service personnel. In many cases, these tests require maximal or near-maximal effort by the subject and thus have not been used for older persons or those at increased risk for cardiovascular disease. Their advantage is that large numbers of subjects can be tested rapidly at low cost. However, to obtain an accurate evaluation, subjects must be willing to exert themselves and know how to set a proper pace.

Muscle endurance

In contrast with cardiorespiratory endurance, muscle endurance is specific to each muscle group. Few tests of muscular endurance for use in the general population are solely endurance measures, however, because most are also tests of muscle strength. Tests of muscular endurance and strength include sit-ups, pushups, the bent-arm hang, and pull-ups. These tests need to be properly administered and may not discriminate well in some populations (eg, pull-ups are not suitable for many populations because a substantial percentage of those tested will have a score of 0). Few laboratory tests of muscle endurance have been developed. Such tests usually involve having the subject perform a series of contractions at a set percentage of maximal strength and at a constant rate until the person can no longer continue at that rate. The total work performed or the test duration is used as a measure of muscle endurance.

Muscle strength

Muscle strength can be measured during performance of either static or dynamic muscle contraction (57). Like muscle endurance, strength is specific to the muscle group, and therefore the testing of one muscle group does not provide accurate information about the strength of other muscle groups (58). Thus, to be effective, strength testing must involve at least several major muscle groups, including the upper body, trunk, and lower body. Standard tests have included the bench press, leg extension, and biceps curl with free weights. The heaviest weight a person can lift one time through the full range of motion is considered the person's maximum strength.



Flexibility

Flexibility is a difficult component to measure accurately and reliably because it is specific to the joint being tested; no one measure provides a satisfactory index of an individual's overall flexibility (59). Field testing of flexibility frequently has been limited to the sit-and-reach test, which is considered a measure of lower back and hamstring flexibility. Other tests have been used to determine the flexibility of the shoulder, hip, knee, and ankle. The criterion method for measuring flexibility in the laboratory is goniometry, which is used to measure the angle of the joint at both extremes in the range of motion (57).

Balance, agility, and coordination

Balance, agility, and coordination are especially important in older persons, who are more prone to fall and as a result suffer fractures because of their reduced bone mineral density. There are no generally accepted standard techniques for measuring balance, agility, and coordination, especially in older persons. Field methods include various "balance stands" (eg, standing on one foot with eyes open and with eyes closed, standing on a narrow block) and "balance walks" on a narrow line or rail (60). In the laboratory, computer-based technology is now being used to evaluate balance measured on an electronic force platform or by analysis of a video recording of the subject walking (61). Agility is usually measured by tests that require rapid changes in body position or changes in direction while walking or running (62). More test development is needed to establish norms for older persons on standardized tests for measuring balance, agility, and coordination.

CONSIDERATIONS IN DESIGNING STUDIES TO EVALUATE THE EFFECTS OF DIETARY SUPPLEMENTS ON PHYSICAL PERFORMANCE

To accurately determine whether a particular dietary supplement significantly benefits physical performance, a scientific evaluation should be performed that includes specific design elements. Many of the claims made for various supplements are based on less-than-rigorous science and thus are not accepted by many in the scientific, medical, nutrition, and exercise communities. At the same time, because the potential benefits of dietary supplements are enticing, supplement providers, coaches, and athletes would like the claims to be true. Becoming more familiar with the design elements that researchers consider essential for a scientifically sound study will ensure that future studies examining the effects of a specific supplement on performance are scientifically rigorous, accurate, reliable, and unbiased.

Placebo control group, blind assignment, and random assignment

Studies examining the effects of dietary supplements on performance must include a placebo control group, in which the athletes are given an inert placebo that looks, tastes, and smells like the dietary supplement given to the athletes in the treatment group. The performance of the placebo control group provides a comparison with that of the treatment group in case variables other than the dietary supplement affect the athletes' performance during the study. For instance, both the treatment and control groups may improve their performance because the weather is no longer hot and humid. However, without the performance of the control group for comparison, one may mistake-

only conclude that the dietary supplement was responsible for the improvement in performance rather than the weather.

In addition, the athletes must be blind to their group assignment; that is, they cannot know whether they are taking the inert placebo or active dietary supplement. The procedure of blind assignment allows for any possible psychological effects on the athletes' performance to be similar in both the treatment and placebo control groups. Athletes may have strong beliefs that the supplement will improve their performance, perhaps because of such claims by a coach, trainer, other athletes, or in promotional material. Athletes with these beliefs may work harder in practice and thus actually improve their performance. Therefore, both the treatment and control groups may improve their performance because they believe the supplement will enhance their performance. Without the blinding of all the athletes and the existence of the placebo control group, one may mistakenly conclude that the dietary supplement was responsible for the improvement in performance, rather than the athletes' belief that the supplement will work. Alternatively, coaches who believe the supplement affects performance may unintentionally push the athletes who are taking the supplement to work harder in practice and thus directly affect the athletes' performance. Thus, both the treatment and control groups may improve their performance because of their coaches' beliefs in the effectiveness of the supplement. If coaches and any other personnel responsible for measuring the performance of the athletes are not blind to the athletes' assignment during the course of the study, one might mistakenly decide that the dietary supplement was responsible for the improvement in performance. Blind assignment may sound difficult or expensive, but it has been easily and inexpensively implemented in many studies evaluating cardiovascular drugs in which the performance of patients on an exercise test is the major outcome.

Studies examining the effects of dietary supplements on performance must also randomly assign athletes to either the treatment or placebo control group. Random assignment distributes any characteristics of the athletes that might influence their performance into the treatment and placebo control groups in approximately the same manner and thus cannot differentially influence the athletes' performance. For instance, in a particular sport, younger athletes may be faster. If all the younger athletes were put in the treatment group, one might mistakenly conclude that the dietary supplement rather than the age of the athletes was responsible for the improvement in performance. Random assignment increases the probability that the younger athletes will be equally distributed between the 2 groups. To use another example, if athletes (or coaches) are allowed to choose whether they want to take the dietary supplement or inert placebo, athletes who believe supplements improve performance may be more likely to select the treatment group, thus biasing the results. It is important to remember that random assignment is designed not only to distribute factors known to influence performance equally between the treatment and placebo control groups but also, and even more important, to equally distribute factors not measured or whose effects on performance are unknown. Random assignment should take place after all the baseline data for an athlete have been collected and the athlete is sure about participating in the study. Study results are substantially weakened if athletes drop out after they have been randomly assigned to a group and before the study is completed (63). For instance, if athletes who improve their performance during the study are more likely to complete the study and come back for a final performance



measurement than athletes who do not improve their performance, one may mistakenly conclude that the study was more successful than it was. In addition, if more athletes drop out of one group than the other (selective dropout), the validity and results of the study can be greatly jeopardized.

Importance of statistical power

To successfully move research about the effects of dietary supplements on the performance of elite athletes from controlled testing in the clinic or laboratory to performance in actual competitions requires studies with adequate statistical power to detect a clinically meaningful (and statistically significant) treatment effect. Statistical power is the probability that the study can detect a statistically significant treatment effect; that is, that it can detect a difference in performance between athletes randomly assigned to receive a dietary supplement and those assigned to receive a placebo, if indeed a treatment effect exists (64). The greater the statistical power, the more likely the study can detect a true treatment effect. The degree of statistical power in a study can be thought of as the strength of a flashlight beam needed to see the size of 2 animals fighting in a backyard. The stronger the flashlight beam, the more easily the animals can be seen. The degree of statistical power in a study is extremely important because it does not affect whether a treatment effect actually exists but rather whether an effect can be detected. Similarly, the strength of the flashlight beam does not affect whether the animals are actually in the backyard but rather how well they can be seen if they are present. Thus, having sufficient statistical power in a study to detect a treatment effect, if it is there, is crucial to evaluating whether dietary supplements improve the performance of elite athletes.

Typically, proposals for well-controlled clinical trials submitted to the National Institutes of Health for funding are designed to achieve 80% or 90% power, that is, an 80% or 90% chance of finding a statistically significant treatment effect if it exists. The degree of statistical power in a study is influenced by 4 factors: sample size, effect size (ie, magnitude of the difference between the treatment and control groups), type of statistical test used, and level of statistical significance (65). This level is often set by convention (eg, $P < 0.05$), but the other 3 factors can be improved in ways that greatly increase the statistical power of a study.

The most common way to achieve sufficient statistical power is to have a large sample size. Unfortunately, a review of published studies to date that examined the effects of dietary supplements on various measures of performance and health revealed that a surprisingly large number of studies can be discarded immediately because of extremely small sample sizes. Because these studies were woefully underpowered, the probability that a treatment effect could have been detected (if it existed) would have been extremely low.

The problem of small sample sizes and inadequate statistical power is particularly damaging when the size of the effect to be detected is very small (ie, a difference of $<5\%$). Randomized trials examining the health benefits of regular physical activity have the goal of detecting a clinically meaningful difference on physical or psychological health measures between the treatment and control groups. For instance, a clinically meaningful effect of moderate-intensity physical activity on the reduction of coronary heart disease mortality as measured in secondary prevention studies would be $\approx 20\%$ (65). Thus, a researcher designing a

prospective secondary prevention study with sufficient statistical power must determine how to achieve an 80% probability of detecting a clinically meaningful 20% reduction in mortality.

If an intervention is safe, inexpensive, and convenient, or if there is no other treatment for a life-threatening disease, a smaller effect might be considered clinically meaningful. In other contexts (eg, competitive sports) as well, a small effect may be considered meaningful. From the perspective of a track coach of Olympic runners, small improvements in performance would be exceptionally valuable and would provide adequate justification for recommending a dietary supplement. This is because the difference in performance between elite runners who win medals and those who finish sixth is extremely small. An analysis of data for men's events averaged across 7 Olympic games (1968–1992) found a difference of just 2.2% (0.22 s) in the 100-m dash between the average time of competitors who won the gold medal (10.03 s) and those who finished sixth (10.25 s). In the 1500-m run, there was a difference of 3.0% (6.6 s) between the gold medalists (3 min, 36.8 s) and sixth-place finishers (3 min, 43.4 s). Similarly, in the men's marathon, there was a difference of 2.3% (180 s) between the gold medal winners (2 h, 12 min, 25 s) and sixth-place finishers (2 h, 15 min, 25 s). From these data, it is apparent that in the Olympics a difference of only 2–3% separates elite male runners who win the gold medals from those who finish in sixth place.

Unfortunately, designing a study to detect a performance difference of only 2–3% with adequate statistical power is difficult because the sample size needed may be prohibitively large. For a hypothetical clinical trial (with a placebo control group and both blind and random assignment) that would examine the effects of a dietary supplement on competitive performance, a standardized effect size must be estimated from pilot data to calculate the sample size needed to achieve a statistical power of 80%. In this example, the pilot data are collected from a small study of 24 1500-m runners who have an average personal best performance of 3 min and 43 s (approximately sixth place in the recent Olympic games), with 12 runners randomly assigned to receive the dietary supplement and 12 assigned to a placebo control group. After 2 mo, the treatment group improves their performance, on average, by 3.3 s, an improvement that is approximately half of the performance difference between gold medal winners and sixth-place finishers in this event (6.6 s). Given that the treatment group improves by 3.3 s (with an SD of change of 10.0 s) and the placebo control group improves its performance by 0.5 s (with a similar SD of change of 10.0 s), the treatment group has a net improvement of 2.8 s ($3.3 - 0.5$).

Based on these pilot data, a standardized effect size is calculated by dividing the difference between the mean change of the treatment and control groups by the pooled SD for the 2 groups ($(3.3 - 0.5)/10.0 = 0.28$). (The pooled SD is a weighted mean of the SD of the change for the 2 groups.) Based on the estimated effect size (0.28), the sample size needed for the hypothetical clinical trial can be determined by using tables or charts that provide sample sizes for a range of effect sizes and that take into account other study design or statistical considerations (64, 66). For instance, shown in **Table 3** is the number of athletes required for each group for different effect sizes assuming a 2-group comparison (eg, dietary supplement compared with placebo control), statistical significance set at $P < 0.05$, a 2-tailed t test, and minimum statistical power of 80% or 90%. As indicated in the table, the hypothetical randomized clinical trial with an estimated



TABLE 3
Sample size per group for different standardized effect sizes

Effect size ¹	Statistical power	
	0.80	0.90
0.20	393	525
0.30	174	234
0.40	98	131
0.50	63	84
0.60	44	58
0.80	25	33
1.00	16	21

¹Effect size = mean of change in treatment group – mean of change in control group/pooled SD of the change. Note: $P \leq 0.05$, 2-tailed, 2 groups.

effect size of 0.28 would require >201 athletes per group to achieve a statistical power of 80%.

Conducting the hypothesized clinical trial with far <200 athletes per group would result in insufficient statistical power, drastically reducing the ability to detect a treatment effect. For instance, if the hypothesized trial were conducted based on the same estimated effect size of 0.28 but with a sample size in each group similar to that typically found in the literature (eg, 12–15), the statistical power would be <15%. Stated another way, if a treatment effect actually exists, the hypothesized trial would have less than a 15% chance of finding a statistically significant difference between the 2 groups.

Increasing the statistical power of a study without increasing the sample size

Although the statistical power of a trial can be increased by increasing the sample size, this strategy can be expensive and can create logistical problems by making staff spend too much time and resources on recruiting and assessing subjects. Alternatively, because dropout greatly decreases the statistical power of the trial, extensive efforts can focus on limiting the number of athletes who drop out of the trial once they are randomly assigned to a group (63).

As noted above, statistical power can also be influenced by the effect size and the type of statistical test used to analyze the data. By improving these factors, the hypothetical clinical trial can achieve sufficient statistical power with a smaller sample size (67). The easiest way to increase the effect size is either to increase the numerator (ie, the difference between the mean change of the treatment and control groups) or to decrease the denominator (ie, the pooled SD of the change, or the variability of the athletes' performance within the groups). For example, administering the most potent dose of the dietary supplement that is still safe to the treatment group will increase the difference between the 2 groups, increase the numerator, and thus increase the effect size (64). In addition, using either an inert placebo or very low dosage in the control group will also increase the difference between the treatment and control groups and thus increase the numerator (64). When a limited number of athletes are available for study recruitment, the statistical power of a trial can be increased by comparing only 2 groups (treatment and placebo control) rather than many treatment groups assigned graduated doses of the supplement.

Several strategies can decrease the denominator, or the variability of the athletes' performance within the groups. For instance, ensuring that the dietary supplement and placebo are

administered uniformly to the athletes will reduce the variability of response within the groups and thus decrease the denominator. Alternatively, using a more reliable measure of performance will decrease the variability in the measurement of change within the groups and thus also decrease the denominator (67). The reliability of a measure can be improved by careful standardization of the measurement procedures among the athletes (64, 67). Finally, statistically adjusting for factors related to the variability in the athletes' performance—such as age, sex, or initial or baseline levels of the performance measures—can also decrease the denominator.


Because statistical power can be influenced by increasing the effect size, the hypothetical clinical trial can be conducted with sufficient statistical power with a smaller sample size. For example, by determining the effect of increasing the potency of the dose of dietary supplement, using an inert placebo control group, and employing a variety of strategies to decrease the variability within the groups in the pilot study, the resulting estimated effect size could be increased from 0.28 to 0.66. Thus, the sample size needed to achieve a level of statistical power of 80% in the hypothetical clinical trial would decrease from >200 athletes per group to 37 athletes per group.

SUMMARY AND RECOMMENDATIONS

Accurate and reliable measurement of physical activity and physical fitness is critical in conducting research designed to evaluate how physical activity influences dietary requirements and whether supplements can enhance physical performance. Methodology for the measurement of physical activity by questionnaires is well developed, and new technologies are being developed and evaluated for assessing body movement or correlates of activity, including accelerometers and doubly labeled water. Laboratory and field methods are available for measuring the various components of physical fitness, with many having the accuracy and reliability to measure many small changes in fitness because of exercise training or dietary supplements. Major limitations of existing research evaluating the effects of dietary supplements on physical or athletic performance have included failure to use a randomized, placebo-controlled double-blind design and inadequate power to establish that differences that are meaningful to coaches and athletes are statistically significant. Research methods need to be adopted that increase the statistical power of dietary supplement studies, including increasing sample size, maximizing treatment effectiveness, selecting appropriate testing procedures (accurate, reliable, and sensitive to change), and enhancing retention of subjects assigned to treatment groups.

Future research should continue to develop measurement methodology for more accurately assessing a person's physical activity profile throughout the day, including a profile of activity intensity and total energy expenditure. Methods are needed that keep subject and investigator burden to a minimum through the use of automated recording and analysis procedures. These methods need to be designed for persons at the high end of the physical activity continuum (such as elite athletes) and those at the low end (such as patients and the very old), because both may benefit by an enhanced understanding of the interactions between dietary supplement use, activity, and physical performance capacity. The emphasis of future research on methods to measure physical fitness should be on procedures to accurately measure changes in performance among persons with a low



performance capacity (patients, obese persons, and the elderly) and on those components of fitness, such as endurance capacity, muscle endurance, and balance, for which standardized testing procedures are not readily available. Efforts should be made to ensure that future research evaluating the effects of dietary supplements on physical performance is appropriately designed, with the statistical power to detect meaningful results. 

REFERENCES

- 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.
- Park RJ. Measurement of physical fitness: a historical perspective. Washington, DC: US Department of Health and Human Services, Public Health Service, 1989:1–35. (Office of Disease Prevention and Health Promotion Monograph Series.)
- President's Council on Physical Fitness and Sports. Exercise programs for adults. Washington, DC: US Government Printing Office, 1965.
- Corbin CB, Lindsey R. Concepts in physical education with laboratories and experiments. 4th ed. Dubuque, IA: Wm C Brown Publishers, 1990.
- Pate RR, Pratt M, Blair SN, 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.
- Bouchard C, Shephard RJ, Stephens T, Sutton JR, McPherson BD. Exercise, fitness and health: the consensus statement. In: Bouchard C, Shephard RJ, Stephens T, Sutton JR, McPherson BD, eds. Exercise, fitness and health. Champaign, IL: Human Kinetic Publishers, 1990:4–28.
- National Center for Health Statistics. Assessing physical fitness and physical activity in population-based surveys. Washington, DC: US Government Printing Office, 1989. (DHHS publication 89-1253.)
- Wilson PFW, Paffenbarger RS, Morris JN, Havlik RJ. Assessment methods for physical activity and physical fitness in population studies; a report of a NHLBI workshop. *Am Heart J* 1986;111:1177–92.
- Stone EF, Sopko G, Haskell WL, Douglas PS, et al, eds. Physical activity and cardiovascular health: special emphasis on women and youth. *Med Sci Sports Exerc* 1992; 24:5191–307.
- Ainsworth BE, Montoye HJ, Leon AS. Methods of assessing physical activity during leisure and work. In: Bouchard C, Shephard RJ, Stephens T, eds. Physical activity, fitness, and health: international proceedings and consensus statement. Champaign, IL: Human Kinetics Publishers, 1994:146–59.
- Morris JN, Heady JA, Raffle PAB, Roberts CG, Parks JA. Coronary heart disease and physical activity of work. *Lancet* 1953;2:1053–7, 1111–20.
- Manson JE, Nathan DM, Krolewki AS, Stampfer MJ, Willet WC, Henkens CH. A prospective study of exercise and incidence of diabetes among US male physicians. *JAMA* 1992;268:63–7.
- Paffenbarger RS Jr, Hyde RT, Wing AL, Lee IM, Jung DL, Kampert JB. The association of changes in physical-activity level and other lifestyle characteristics with mortality among men. *N Engl J Med* 1993;328:538–45.
- Jacobs DR, Ainsworth BE, Hartman TJ, Leon AS. A simultaneous evaluation of ten commonly used physical activity questionnaires. *Med Sci Sports Exerc* 1993;25:81–91.
- Montoye HJ. Measuring physical activity and energy expenditure. Champaign, IL: Human Kinetics, 1995.
- Caspersen CJ. Physical activity epidemiology: concepts, methods, and applications to exercise science. *Exerc Sport Sci Rev* 1989;17:423–73.
- Kannel WB, Wilson PWF, Blair SN. Epidemiological assessment of the role of physical activity and fitness in development of cardiovascular disease. *Am Heart J* 1985;109:876–85.
- Powell KE, Paffenbarger RS. Workshop on Epidemiologic and Public Health Aspects of Physical Activity and Exercise: a summary. *Public Health Rep* 1985;100:118–26.
- LaPorte RE, Montoye HJ, Caspersen CJ. Assessment of physical activity in epidemiologic research: problems and prospects. *Public Health Rep* 1985;100:131–46.
- Montoye HJ, Taylor HL. Measurement of physical activity in population studies: a review. *Hum Biol* 1984;56:195–216.
- King AC, Haskell WL, Taylor CB, Kraemer HC, DeBusk RF. Groupvs home-based exercise training in healthy older men and women. A community-based clinical trial. *JAMA* 1991;166:1535–42.
- Baranoski T. Methodological issues in self-report of health behaviors. *J Sch Health* 1985;55:179–82.
- Blair SN, Dowda M, Pate RR, et al. Reliability of long-term recall of participation in physical activity by middle-aged men and women. *Am J Epidemiol* 1991;133:266–75.
- Kriska AM, Sandler RB, Cauley JA, LaPorte RE, Hom DL, Pambianco, G. The assessment of historical physical activity and its relation to adult bone parameters. *Am J Epidemiol* 1988;127:1053–63.
- Montoye HJ, Cunningham DA, Welch HG, Epstein FH. Laboratory methods of assessing metabolic capacity in a large epidemiologic study. *Am J Epidemiol* 1970;91:38–47.
- Taylor HL, Jacobs DR Jr, Schucker B, Knudsen J, Leon AS, Debacker G. A questionnaire for the assessment of leisure time physical activities. *J Chronic Dis* 1978;31:741–55.
- Bloom B. Current estimates from the National Health Interview, United States, 1981. *Vital Health Stat* 10 1982;141. (US DHHS publication PHS 83-1569.)
- Wasburn RA, Adams LL, Haile GT. Physical activity assessment for epidemiologic research: the utility of two simplified approaches. *Prev Med* 1987;16:636–46.
- Caspersen CJ, Pollard RA. Validity of global self-reports of physical activity in epidemiology. *CVD Epidemiol Newsletter* 1988;43:15.
- Dipietro L, Caspersen CJ, Ostfeld AM, Nadel ER. A survey for assessing physical activity among older adults. *Med Sci Sports Exerc* 1993;25:628–42.
- Voorrips LE, Ravelli AC, Dongelmans CA, Deurenberg P, Van Staveren WA. A physical activity questionnaire for the elderly. *Med Sci Sports Exerc* 1991;23:974–9.
- Washburn RA, Smith KW, Jette AM, Janney CA. The physical activity scale for the elderly (PHASE): development and evaluation. *J Clin Epidemiol* 1993;46:153–61.
- Sallis JF, Condon SA, Goggin KJ, Kolody B, Alearaz JE. The development of self-administered physical activity surveys for 4th grade students. *Res Q Exerc Sport* 1993;64:25–31.
- 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.
- Gilliam TB, Freedson PS, Geenen DL, Shahraray B. Physical activity patterns determined by heart-rate monitoring in 6–7 year-old children. *Med Sci Sports Exerc* 1981;13:65–7.
- Sallis JF, Buono MJ, Roby JJ, Carlson D, Nelson JA. The Caltrac accelerometer as a physical activity monitor for school-age children. *Med Sci Sports Exerc* 1990;22:698–703.
- Freedson PS. Field monitoring of physical activity in children. *Pediatr Exerc Sci* 1989;1:8–18.
- Kashiwazaki H, Inaoka T, Suzui T, Kondo Y. Correlations of pedometer readings with energy expenditure in workers during free-living daily activities. *Eur J Appl Physiol* 1986;54:585–90.
- Washburn RA, Janney CA, Fenster JR. The validity of objective physical activity monitoring in older individuals. *Res Q Exerc Sport* 1990;61:114–7.
- LaPorte RE, Black-Sandler R, Cauley JA, Link M, Bayles C, Marks B. The assessment of physical activity in older women: analysis of the interrelationship and reliability of activity monitoring, activity surveys, and caloric intake. *J Gerontol* 1983;34:394–7.

41. Wong TC, Webster JG, Montoye HJ, Washburn R. Portable accelerometer device for measuring human energy expenditure. *IEE Trans Biomed Eng* 1981;28:467-71.
42. Taylor CB, Kraemer HC, Bragg DA, et al. A new system for long-term recording and processing of heart rate and physical activity in outpatients. *Comput Biomed Res* 1982;15:L7-17.
43. Balogun J, Amusa LO, Onyewadume IU. Factors affecting Caltrac and Calcount accelerometer output. *Phys Ther* 1988;68:1500-4.
44. Haskell WL, Yee MC, Evans A, Irby PJ. Simultaneous measurement of heart rate and body motion to quantitate physical activity. *Med Sci Sports Exerc* 1993;25:109-15.
45. Klesges LM, Klesges RC. The assessment of children's physical activity: a comparison of methods. *Med Sci Sports Exerc* 1987;19:511-7.
46. Rogers F, Juneau M, Taylor CB, et al. Assessment by a micro-processor of adherence to home-based moderate-intensity exercise training in healthy, sedentary middle-aged men and women. *Am J Cardiol* 1987;60:71-5.
47. Richardson MT, Leon AS, Jacobs DR Jr, Ainsworth BE, Serfass R. Ability of the Caltrac accelerometer to assess daily physical activity levels. *J Cardiopulm Rehabil* 1995;15:107-13.
48. Luke A, Maki KC, Barkey N, Cooper R, McGee D. Simultaneous monitoring of heart rate and motion to assess energy expenditure. *Med Sci Sports Exerc* 1997;29:144-8.
49. Klein PD, James WP, Wong WW, et al. Calorimetric validation of the doubly-labelled water method for determination of energy expenditure in man. *Hum Nutr Clin Nutr* 1984;38:95-106.
50. Westerterp KR, Brouns F, Saris WH, ten Hoor F. Comparison of doubly labeled water with respirometry at low- and high-activity levels. *J Appl Physiol* 1988;65:53-6.
51. Malina R, Bouchard C. Genetic considerations in physical fitness. In: Drury TF, ed. *Assessing physical fitness and physical activity in population based surveys*. Washington, DC: National Center for Health Statistics, 1989. (US DHHS publication PHS 89-1253.)
52. Mitchell JH, Blomqvist G. Maximal oxygen uptake. *N Engl J Med* 1971;284:1018-22.
53. Siconolfi SF, Cullinane EM, Carleton RA, Thompson PD. Assessing VO_2max in epidemiologic studies: modification of the Astrand-Rhyming test. *Med Sci Sports Exerc* 1982;14:335-8.
54. Blair SN, Kohl HW 3rd, Paffenbarger RS Jr, Clark DG, Cooper KH, Gibbons LW. Physical fitness and all-cause mortality: a prospective study in healthy men and women. *JAMA* 1989;262:2395-401.
55. Sidney S, Haskell WL, Crow R, et al. Symptom-limited graded treadmill exercise testing in young adults in the CARDIA study. *Med Sci Sports Exerc* 1992;24:177-83.
56. Cooper KH. A means of assessing maximal oxygen uptake. *JAMA* 1968;203:201-4.
57. Wilmore JH. Design issues and alternatives in assessing physical fitness among apparently healthy adults in a health examination survey of the general population. In: *Assessing physical fitness and physical activity in population-based surveys*. Washington, DC: National Center for Health Statistics, 1989. (US DHHS publication 89-1253.)
58. Clarke HH. Toward a better understanding of muscular strength. *Physical Fitness Research Digest* 1973;3:1-20.
59. Harris MC. A factor analytic study of flexibility. *Res Q* 1969;40:62-70.
60. Tse SK, Bailey DM. T'ai chi and postural control in the well elderly. *Am J Occup Ther* 1992;46:295-300.
61. Lehmann JF, Boswell S, Price R, et al. Quantitative evaluation of sway as an indicator of functional balance in post-traumatic brain injury. *Arch Phys Med Rehabil* 1990;71:955-62.
62. Cureton TK. *Physical fitness workbook*. 1st ed. St Louis: Mosby, 1943:1-188.
63. Ribisl KM, Walton MA, Mowbray CT, Luke DA, Davidson WS, Bootsmliller BJ. Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: review and recommendations. *Eval Program Plann* 1996;19:1-25.
64. Lipsey MW. *Design sensitivity: statistical power for experimental research*. Newbury Park, CA: Sage Publications, 1990.
65. Oldridge NB, Guyatt GH, Fisher ME, Rimm AA. Cardiac rehabilitation after myocardial infarction: combined experience of randomized clinical trials. *JAMA* 1988;260:945-50.
66. Kraemer HC, Thiemann, S. How many subjects? Statistical power analysis in research. Newbury Park, CA: Sage Publications, 1987.
67. Kraemer HC. To increase power in randomized clinical trials without increasing sample size. *Psychopharmacol Bull* 1991;27:217-24.

