Seeing Is Believing: Observing Physical Activity and Its Contexts

Thomas L. McKenzie

Direct (systematic) observation has been a mainstay of my research for over three decades. I believe it is an important tool for assessing physical activity, because it can simultaneously provide contextually rich data on the setting in which the activity occurs. It is particularly useful for those interested in using ecological and cognitive-behavioral approaches to examine how physical and social environments influence physical activity, and it is currently the only method for assessing group physical activity in open environments. Because physical activity researchers use the method infrequently, this paper provides an overview of systematic observation as it applies to studying physical activity.

Key words: environment, exercise, observational research, research methods

Being selected to deliver the 2009 C. H. McCloy Research Lecture is a highlight of my academic career. McCloy was a distinguished scholar whose primary area of expertise was measurement, particularly of physical fitness and motor skills. Context does matter, and I will address the measurement of physical activity—an aspect of little concern to kinesiology and the health professions during McCloy’s distinguished career at the University of Iowa from 1930 to 1959. Specifically, I will focus on the use of systematic, direct observation methodology to measure physical activity and the many contexts in which it occurs.

Systematic observation is a procedure or method for generating data, not a single technique or instrument. Early in my career I used direct observation primarily to study behaviors in sports (e.g., McKenzie & Rushall, 1974) and physical education (e.g., McKenzie, 1981). With the increasing global concern for sedentary living, I began to focus on investigating health promoting physical activity in diverse locations, including homes, schools (e.g., physical education, recess, and before and after school), and more recently in park and recreation settings.

Numerous observation systems have been developed, and in an earlier paper I compared nine instruments for studying children’s physical activity on 11 different characteristics (McKenzie, 2002). All nine had been carefully constructed, validated using energy expenditure estimations, field-tested with children, had high reliabilities, and were generalizable for use in studying physical activity as it relates to health. All nine used sampling procedures to estimate the time children spent in various activity categories; eight targeted individuals, and one targeted groups. No instruments validated for studying adult physical activity levels were available at that time.

The purpose of the current paper is not to present another review of observation systems but to serve as a tutorial on using systematic observation to investigate physical activity and associated variables. It includes background information on the science of systematic observation and...
refers to our investigations to illustrate the technical and practical considerations often encountered when using direct observation to assess physical activity in natural settings. I refer to “our” work, because systematic observation research is never done alone (e.g., need for others to assist with reliability measures), and over the years I have had numerous collaborators. Table 1 summarizes five observation systems that we developed, evaluated, and made available to the public. While all use similar physical activity codes, each has a different research focus. Three focus on individuals’ physical activity in specific environments (Behaviors of Eating and Activity for Children’s Health: Evaluation System [BEACHES] in homes, System for Observing Fitness Instruction Time [SOFIT] in physical education classes, and System for Observing Children’s Activity and Relationships During Play [SOCARP] in free play settings), and two focus on assessing group physical activity in open environments (System for Observing Play and Leisure in Youth [SOPLAY] in school play areas and System for Observing Play and Active Recreation in Communities [SOPARC] in community and park settings). 1

Table 1. Sample Physical Activity Observation Systems

The five systems outlined here are based on the same five physical activity level codes. These levels have been validated by heart rate monitoring, oxygen consumption, accelerometer, and pedometer. Validation studies have involved children from preschool to 12th grade and included those with developmental delays.

1. SOFIT: System for Observing Fitness Instruction Time
Provides simultaneous data on (a) student activity levels, (b) the lesson context in which they occur (i.e., how lesson content is delivered, including time for fitness, skill drills, game play, knowledge, and management), and (c) teacher interactions relative to promoting physical activity and fitness. Teacher gender, class gender composition, lesson location, and number of students in class are also recorded.

2. SOPLAY: System for Observing Play and Leisure in Youth
Provides data on the number of participants and their physical activity levels during play and leisure opportunities in targeted areas. Separate scans are made for girls and boys, and simultaneous entries are made for area contextual characteristics including their accessibility, usability, and whether or not supervision, organized activities, and equipment are provided. The predominant type of activity engaged in by area users is also recorded (e.g., basketball, dance).

3. SOPARC: System for Observing Play and Active Recreation in Communities
Provides data on the number of participants and their physical activity levels in park and recreation settings. Simultaneously codes contextual characteristics of areas including their accessibility, usability, and whether or not supervision, organized activities, and equipment are provided. Expands SOPLAY by making separate scans for race/ethnicity and age groupings.

4. SOCARP: System for Observing Children’s Activity and Relationships During Play
Purpose. Provides data on children’s physical activity levels on the playground while simultaneously assessing the contextual variables of social group size, activity type, and pro- and anti-social interactions with peers.

5. BEACHES: Behaviors of Eating and Activity for Children’s Health: Evaluation System
Provides data on children’s at home physical activity and sedentary behaviors and selected environmental (social and physical) variables that may influence these events. Includes data on child location, presence of others, prompts and consequences for physical activity and sedentary behavior, food ingesting, and media viewing.

Background and Rationale for Using Systematic Observation to Assess Physical Activity

Behavioral observation has played an important role in sport, physical education, and dance, for by observing directly most of the judgments teachers, coaches, and officials made were on the quantity and quality of movements and the outcomes of games and performances. Additionally, direct (and systematic) observation has had a long history in studying human behavior in natural settings (Ciminero, Calhoun, & Adams, 1977; Cooper, Heron, & Heward, 2007; Hartman, 1982), including sport and physical education environments (Darst, Zakrjesk, & Mancini, 1989). Although it has been a common and essential method for studying behavior in a wide range of applied disciplines (e.g., psychology, sociology, zoology), using direct observation in assessing physical activity has historically been viewed as labor intensive and tedious (Montoye, Kemper, Saris, & Washburn, 1996) and is still often overlooked as a viable method (e.g., Corder, Ekeland, Steele, Wareham, & Brage, 2008).
Physical activity is the process of engaging in bodily movement that results in energy expenditure. Its beneficial effects on many bodily systems in adults have been known for some time, and its effects on children are becoming more clear (U.S. Department of Health and Human Services [USDHHS], 2008a), but only recently have the first national recommendations for physical activity and health promotion and disease prevention been released (USDHHS, 2008b). As a concept, physical activity is defined by transitory and multidimensional behaviors (e.g., frequency, duration, intensity, mode), thus, making it challenging to assess. Numerous methods to measure physical activity are used, including questionnaires, accelerometry, heart rate monitoring, double-labeled water, and direct observation. Each method has strengths and limitations (Gorder et al., 2008; Welk, 2002). Because of ease of use, self-reports on questionnaires is the most common way researchers assess physical activity. This method, however, has substantial limitations related to recall bias and time estimations—especially with children. Using questionnaires and interviews with children below the fourth-grade level is particularly problematic, because most children are inept at recalling the frequency, duration, and intensity of their physical activity.

Direct observation exceeds other physical activity measures in providing contextually rich data on the environment, including social (e.g., verbal interactions) and physical (e.g., presence of structures or equipment) variables. Thus, observation is a particularly useful method for using cognitive-behavioral and ecological approaches to understand how physical activity occurrence is influenced by its contexts (Sallis, Owen, & Fisher, 2008). In addition, because technological advances permit observational data to be entered, stored, and analyzed by hand-held computers, the methodology is now much more appealing than it was previously.

Direct observation can be used to generate formative, process, and outcome data, and its strength rests on its ability to measure physical activity as well as its capacity to identify the type of activity and when, where, and with whom it occurs. As a direct measure of behavior in the natural setting, it minimizes inferences being made concerning the actions. Observation systems can be flexible in their design, permitting researchers and practitioners to focus on different interest variables; because assessors do not interact with those they observe, there is little participant burden. Participants in observational studies are not asked to recall, wear, or inject anything, and in open settings (e.g., parks) those being observed may not even know their physical activity is being measured.

Systematic observation is especially important in assessing health promoting physical activity, which Sallis (2009) described as being place-dependent. Physical activity is always contextual and strongly influenced by proximal social and physical environmental factors. Health researchers are interested in assessing a wide variety of environmental (built and natural) variables that affect physical activity, whether it occurs in homes, schools, parks, recreation centers, or work settings.

A special attribute of systematic observation is that it provides researchers and practitioners with a practical tool to assess group physical activity in open environments, such as playgrounds, beaches, and parks (McKenzie, Cohen, Sehgal, Williamson, & Golinelli, 2006; McKenzie, Marshall, Sallis, & Conway, 2000). Other methods, such as heart rate monitoring, accelerometers, and pedometers, measure an individual’s physical activity only and do not provide contextual information on that person’s location. In open environments, however, the number of participants and their characteristics (e.g., age, gender, race/ethnicity) and physical activity levels change frequently, as do the characteristics of the setting. Subsequently, we developed SOPLAY and SOPARC to assess peoples’ characteristics (e.g., activity levels, age, gender, race/ethnicity) as well as modifiable factors in the environment, including whether the activity area is accessible, usable, equipped, organized, and supervised.

Assessment using direct observation has its limitations. Events under study must be observable and codeable, and substantial time and effort are needed to validate instruments. Data collection can also be expensive in terms of observer time, whether observations are done in real time or from video recordings. This limits the number of behavior samples that can be taken and brings into question whether data from those samples are adequate for that setting or generalizable (e.g., to other times of the day or other locations). Observers must be properly trained to be objective and nonjudgmental, and steps must be taken to ensure they maintain their skills over time. Additionally, observations cannot be made in certain environments (e.g., in bathrooms during home observations), and there is the possibility people may behave differently when an observer is present (i.e., reactivity).

**Observing Physical Activity: Measurement Considerations**

**Validity**

Validity refers to a measure assessing what it purports to measure. Because systematic observation is a direct measure of behavior, it requires little inference or interpretation and, subsequently, has high internal validity, sometimes referred to as face validity. As a result, it has been used as a criterion for validating other physical activity measures, including pedometers and accelerometers (Finn & Specker, 2000; McClain, Abraham, Brusseau, & Tudor-Locke, 2008; Scruggs, 2007).
In considering an observation system's validity, it is important to acknowledge that physical activity and energy expenditure are substantially different. Physical activity involves body movement and can be assessed directly through observation methodology. On the other hand, energy expenditure from moving can only be inferred. The movements of a lean child and an obese child may be observed and recorded accurately (e.g., both children walked continuously for 15 min), resulting in useful information, particularly to teachers and parents. The energy expenditure from these similar behaviors, however, would be substantially different, and further calibrations would be needed to ensure the observed data produced a valid estimation of energy expenditure. Because many health researchers are interested in caloric expenditure, most published instruments for observing physical activity have undergone external validations using some measure of energy expenditure estimation (McKenzie, 2002). Figure 1, for example, illustrates the estimated energy expenditure scores derived from one study (McKenzie et al., 1991) in which children wore heart rate monitors while being assessed in the five posture codes used in BEACHES, SOFTI, and SOCARP. Because the estimated energy costs for lying down, sitting, and standing are so similar (McKenzie et al., 1991), the three posture codes in SOPLAY and SOPARC were collapsed into a “sedentary” category for ease of recording.

While external validation is especially important for investigators interested in energy costs, it may not be a priority for practitioners and some physical activity researchers. There are substantial time and financial costs associated with validating physical activity observation codes, and beginning investigators might consider selecting among already validated measures rather than developing one on their own. Most of the published instruments can be modified to permit investigators to modify the contextual codes while keeping the validated physical activity codes intact.

Reliability

Reliability refers to consistency or a measure’s repeatability and is one way to assess the collected data’s quality. In observational research, reliability usually refers to the level of agreement among trained, independent observers and is expressed as an index (typically % agreement) of the degree of correspondence between codes entered in two or more data files. While reliability typically refers to comparing the results obtained by different observers coding the same participant at the same time (i.e., interobserver reliability), it may also refer to the degree of correspondence between the same observer’s scores after coding the same video more than once (i.e., intraobserver reliability).

Different statistics are used in reporting reliability measures, including interval-by-interval (I-I), intraclass correlation, and Kappa. I-I agreement scores are most commonly reported and are particularly useful in assessing the consistency of assessors both during observer training and in the field. By examining individual observation intervals immediately during training, specific problem areas can be identified and subsequently rectified. Reliability scores vary with the complexity of the observation system and the method used to calculate agreement. The Kappa statistic takes chance agreement into consideration, resulting in Kappa scores being somewhat lower than I-I scores. While published observation systems may have been shown to be used reliably, it does not mean a new study using different observers produces reliable data. Thus, reviewers of observational research papers usually demand that reliability scores be reported for each study. While there are many different ways to compute reliabilities (e.g., by individual item, variable, or summary score), those reported in a paper should be for the variables specifically addressed, not for a general score (e.g., a mean interobserver score for all categories combined).

In addition to assessing reliabilities in the field, it is important to implement procedures to assess observer consistency over time and, thus, detect and reduce “instrument decay” or “observer drift.” This is done over the course of a study by having observers periodically score precoded “gold-standard” videos. Any reductions in reliability scores over time would indicate a slippage or a drift from the original coding definitions and be a stimulus for additional training.

Adequacy of the Sample and Time Frame for Observations

Common questions in systematic observation research frequently relate to how much data or how many

---

**Estimated Energy Expenditure**

<table>
<thead>
<tr>
<th>Kcal/kg/min</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
</tr>
</tbody>
</table>

**Figure 1.** Estimated energy expenditure rates of observational categories as determined by heart rate monitoring.
observations are needed and, thus, to determine the measurements needed to correctly classify physical activity for generality purposes and for assessing power calculations to estimate intervention effects. With the intensive nature of direct observation, it is also important to consider the costs of data collection and the burden placed on observers (e.g., how many variables they can master, how long can they observe accurately before becoming fatigued).

With physical activity being so highly variable, decisions related to how many, how long, and when observations should take place must be determined for each study. Decisions about how much data are needed also depend on the research question and the setting (e.g., physical education lessons, recess, before school, at home with parents present, community parks). To ensure the behaviors are captured within a particular time period, researchers need to visit target environment(s) and collect formative data before settling on a particular observation schedule. Thus, decisions must be made on target location, season, weather, temperature, and time of day. In general, the precise number of observations is unknown, because the number depends on the forementioned factors and others, such as the participants’ age, the physical activity measure used (i.e., observation system), and whether a specific activity level or activity summary score is used to make the determination (Levin, McKenzie, Hussey, Kelder, & Lytle, 2001).

A few studies have been conducted on the number and times of observations. With young children, minute-to-minute variability in activity levels is extremely high, so it is important to measure often (Bailey et al., 1995). Klesges et al. (1984) suggested that at least four observation sessions are needed to adequately estimate the physical activity of young children at home. Along this line, Baranowski, Thompson, Durant, Baranowski, and Puhl (1993) indicated that gender, month, and location accounted for 75% of the variability in the overall activity levels of 3- and 4-year-old children.

In another study using SOFIT, regression models indicated that elementary schools, schools by semester, and weeks explained over 33% of the variability in physical activity provided during physical education lessons (Levin et al., 2001). The magnitude of variation was greater for vigorous physical activity than for either moderate-to-vigorous physical activity or estimated energy expenditure. Variability in activity levels during physical education lessons results from numerous factors, such as lesson goals, content and placement within an instructional unit, teacher behavior, available equipment and facilities (e.g., size of indoor space), and individual differences among children. Each factor needs to be considered when determining the number of observations needed for studying physical education settings.

More recently, we tested SOPARC’s ability to estimate the total number and characteristics of people using a neighborhood park during a week. To determine the most advantageous scheme for accurately estimating park physical activity and use, we observed people in 10 parks in 5 cities across the United States for 14 hr on 14 different days during two seasons. In general, the results indicated that a schedule of observations limited to 4 days per week with four 1-hr observation periods per day was sufficient to estimate park use and park-user characteristics (i.e., user gender, physical activity levels, and age and ethnicity/race groupings; McKenzie et al., 2009). Subsequently, in our newer park studies we reduced the observation schedule from 7 to 4 days a week and are using the savings (time and personnel costs) to add more parks to the sample.

### Technological Considerations in Systematic Observation

Most current systems using direct observation to assess physical activity have their roots in applied behavior analysis (Cooper et al., 2007), which is based B. F. Skinner’s (1953) work in studying operant behavior. Thus, there is interest not only in assessing physical activity behavior directly but also in the simultaneous recording of environmental factors that may be related to the occurrence of that behavior. These potentially determining conditions include physical (e.g., accessibility, equipment availability) and social antecedents (e.g., prompts) and consequences (e.g., praise for activity engagement).

#### Behavior Categories: Classes of Observed Physical Activity

In observational research, the behaviors of interest and their associated characteristics are arranged into groups of mutually exclusive behavior categories called classes (Cooper et al., 2007). Within a class, only one state can occur (and be recorded) at a time. Physical activity levels, location of the target participant, and social interactions are examples of independently scored classes. The five observation systems in Table 1 are based on physical activity being classified into five mutually exclusive categories or codes. The three systems that focus on individuals (BEACHES, SOFIT, SOCARP) code all physical activity into five levels; four are associate with posture (i.e., lying down, sitting, standing, walking), and one (vigorous) is coded when the targeted person’s energy expenditure is greater than an ordinary walk.

For recording ease, lying down, sitting, and standing are collapsed into “sedentary” when using the two systems designed to code groups of people (i.e., SOPARC). In contrast, Bailey’s system (Bailey et al., 1995) has 42 categories—14 posture codes with 3 intensity levels each. More codes and code levels (called subcodes) allow...
greater precision but come at a cost. Additional codes increase the demands of observer training; observing in the field is more tedious, observers become fatigued more quickly, and reliability scores are lowered. The increased precision may be valuable if the codes are translated into energy expenditure scores, but using more codes may not be necessary to answer a particular research question or to be of value to practitioners.

Observation Recording Techniques

Systematic observation uses several data recording techniques, and these have been incorporated into instruments or systems to study physical activity in specific locations or different theoretical backgrounds. Systems, such as those outlined in Table 1, are designed for a specific purpose and include a definition of the behavioral categories of interest, protocols for use (e.g., pacing observations and data entry), and coding conventions (i.e., written or video interpretations of common scenarios). Within a system, various recording techniques may be used to generate data, but they primarily measure the frequency of a behavior (number of times) and its duration (length). Because it is not possible to measure the frequency or duration of all behaviors, sampling is often used.

Sampling Methods. In observational studies, a sampling method specifies which participants to watch, when to watch them, and how to record their behavior. Continuously observing a single participant and coding the onset and ending of each behavior of interest is the most informative and preferred procedure, but this method often is not feasible when multiple behaviors or events are being documented (see Keating, Kulina, & Silverman, 1999). When several behaviors are recorded simultaneously (e.g., physical activity plus prompts and reinforcement for activity), time sampling involving interval recording is often required. In this case, observation periods are typically divided into short observe-record segments. For example, an assessor might observe a child for 10 s and use the next 10 s to enter data. The result, in this case, would be entering three data codes for the child’s physical activity and each associated event during each minute. Observation and recording intervals are typically the same length and usually range from 3 to 35 s.

Several sampling techniques are available, including momentary time (i.e., instantaneous or scan sampling), partial time (recording the event if it occurs at any time during the observe interval), and whole interval (recording the event only if it persists throughout the entire interval). Each method has its advantages and limitations (Cooper et al., 2007), but space does not permit a detailed assessment of each in this paper.

Observation systems often use a combination of sampling techniques. In SOFIT, for example, alternating 10 s observe and 10 s record intervals are typically used for the three main categories (student activity levels, lesson context, and instructor behavior). The physical activity behavior (a continuous event) of a targeted individual is coded at the end of each observe interval using momentary time sampling, as is the context that describes how physical education is delivered. Meanwhile, instructor behaviors of interest, such as physical activity prompting, are not continuous events and, thus, cannot be captured well using momentary time sampling. Subsequently, in SOFIT instructor behavior is coded using partial-interval recording, and physical activity prompting is scored when it occurs at any time during an observation interval.

Duration recording is the preferred method for assessing physical activity using SOFIT and other instruments, because it can provide both the frequency and duration of individual events. Nonetheless, the viability of using momentary time sampling in SOFIT to estimate different physical activity level durations has been demonstrated (Heath, Coleman, Lensegrav, & Fallon, 2006; McNamee & van der Mars, 2005).

Observation Pacing

For time sampling, observers must be prompted when to observe and when to enter codes. Stopwatches initially provided this function, but using them required observers to look away from the viewing area. Audiotape players with prerecorded signals to initiate and end recording replaced stopwatches, but these are outdated; currently the audio signaling is done by MP3s, IPODS, and I-phones. Observers wear a single ear piece so they can hear the pacing signal privately while listening to other sounds in the observational setting. During reliability assessments, two independent observers simultaneously hear the pacing signals by using a “Y” adapter and two separate ear pieces. When data are entered directly into a computer, its internal clock can pace the intervals through either audible signals or visual displays on the screen.

Commercial Hardware and Software

The strengths and limitations of several computer packages for collecting, managing, analyzing, and presenting observational data have been reviewed (Kahng & Iwata, 1998), and more sophisticated systems are regularly becoming available. Several, including the Observer (Prospect Sales & Marketing, Inc., Lilburn, GA) and The Observer (Noldus Information Technology, Leesburg, VA), are comprehensive systems that allow users the flexibility to identify their variables and codes.

Most software packages permit data entry directly into computers, and some allow events to be coded directly from videos and digital media files. Entering data directly into a computer reduces error and speeds analyses and reporting. Handheld computers and PDAs are also
Observer Training and Maintenance

Substantial initial observer training is required, and it is important to monitor observers in the field and retrain them if necessary, because skills may deteriorate or drift over time. The following steps are common in developing and maintaining observer accuracy:

1. Orientation of data collectors to the science of systematic observation
2. Study of the observation manuals and protocols
3. Memorization of response categories
4. Direct practice using modeling and video segments
5. Assessments and feedback using gold-standard video segments
6. Field practice in the observational or similar setting while receiving immediate feedback
7. Field practice and reliability assessments with a certified assessor (before data collection)
8. During data collection periods, include monitoring assessments (i.e., reliabilities) and booster training to detect/prevent observer drift.

Using Video for Observer Training and Maintenance

Numerous video samples are needed for training, certification, and reassessment, particularly in longitudinal studies. For example, SOFIT was used to assess the maintenance effects of a physical education intervention initiated in 96 elementary schools 9 years previously (McKenzie et al., 2003). It would not have been possible to make valid comparisons between data sets collected 9 years apart unless all observers had been trained, certified, and revalidated using the same videotapes and protocol.

In preparing videos, it is important to include samples of all study variables (including all physical activity levels advantageous for collecting physical activity data, because they permit greater mobility in the field. Their small size, low weight, and long battery life often make them preferable to laptops for data collection. Nonetheless, caution is advised because some models have tiny keyboards that increase entry error and some have screens that cannot be seen in the bright sunlight.

Once data are collected, they can be transferred to a stationary computer for further analysis, and commercial systems permit sophisticated data analyses that were not previously possible. For example, during recording one can log data by typing predefined key codes or clicking items on a screen with a mouse or pen stylus. Meanwhile, the computer can automatically add a time stamp to each event and write it to a data file. In duration recording, one can score behaviors (e.g., walking) by pressing a key at the start of the behavior and automatically stop by pressing a key representing another behavior (e.g., walking). Data analysis now can include time-event tables and plots and generate reports with statistics on frequencies and durations, the sequential structure of the process, and the co-occurrence of events. Lag sequential analyses, which examine how often certain behaviors are preceded or followed by other events, are also possible. For example, it is possible to assess how often a parental activity prompt is followed by a child’s actual increase in physical activity and how long it took for that behavior to occur. The software can then translate the information into probability quotients. Regardless of the observation system selected, how often it is used, and how complex the analysis, the extent to which recorded data reflect the actual behavior occurrence largely depends on the skills of human observers.
and contextual categories). For example, BEACHES was originally designed to assess family influences on children’s physical activity and eating (McKenzie et al., 1991). Therefore, during training it was necessary to have video segments that illustrated physical activity levels in various locations, eating behavior, prompts and consequences for physical activity, sedentary behavior, and children’s reactions to the prompts and consequences they received.

Training and assessment videos should be shot in the targeted observation environment or one similar to it. A study being conducted in numerous schools should have video samples from a variety of schools, not just one. Authentic examples of low-occurring variables may be difficult to capture, so it may be necessary to fabricate some video segments to ensure observers can be trained and tested on all constructs.

Shooting video footage that is useful for editing and coding requires careful planning. If children are involved, appropriate school, child, and parent clearance must be obtained before videos are made. People with characteristics (e.g., age, gender, ethnicity) that look and behave like the participants being studied should be on the videos, and they should show people being physically active in various ways. A video of a child “just sitting” during an entire sequence makes coding easy and enables high reliability scores, but it is not useful for training observers.

To facilitate coding, the target participant must be clearly visible on the video (e.g., wearing a brightly colored top or bottom). Zooming in is often necessary to locate the person in a crowd, but there should also be sufficient background to enable observers to accurately record contextual categories, such as the location. If verbal interactions are to be coded, a remote wireless microphone is needed. To avoid distractions, the camera should be mounted on a tripod and panning and tilting should be done slowly and smoothly.

Video technology has improved greatly, and digital video provides high quality images with little loss during editing. Editing software and large computer storage is now commonplace, but amateurs can still be challenged by lack of editing skills. Developing top quality videos still requires consultation with an experienced producer and access to professional editing equipment. This was done in producing DVDs for training observers to use SOPLAY/SOPARC and SOFT. These two DVDs were made so that observers across the country and elsewhere could be trained in a similar manner and compare data from different projects. Both DVDs have the following segments: introduction to the system, video samples of all variables, opportunities to code video segments while receiving immediate feedback (i.e., correct responses appear on the screen after a short delay), and assessment segments in which the appropriate codes (i.e., gold-standard responses) are available only to project directors. The utility of standardized video segments was evidenced during the Child and Adolescent Trial for Cardiovascular Health (McKenzie et al., 1996, 2001, 2003) and Trial of Activity for Adolescent Girls projects (McKenzie, Catellier, et al., 2006) which used SOFIT to assess the physical education interventions. Prior to data collection in different U.S. states, SOFT coordinators and lead observers from each site met for training and certification at a central location. These lead observers then used a training-of-the-trainers model to train additional observers at their local site. The new observers were tested on novel videos that were scored by the project head trainer, and throughout data collection all observers were required to periodically code additional lesson videos to ensure their skills were maintained. Observers falling below standard were retrained and required to code additional videos.

Using Video for Data Collection

In addition to training observers and monitoring their accuracy, videos can be used as a data source. For example, one of the first studies using systematic observation to assess physical activity involved the motion picture analyses of obese and nonobese girls (Bullen, Reed, & Mayer, 1964). While that technology has been replaced by small, portable, low-cost video systems, videos can provide permanent samples that many different observers can view repeatedly and in environments removed from the natural setting. Videos are particularly useful when observation codes are complex, the activity setting is fixed (e.g., a single gymnasium), and only a few participants are involved, such as when validating another physical activity measurement technique.

There are advantages to using video technology to obtain accurate data, but researchers need to consider the disadvantages of this methodology over observing “live,” including time demands, inconvenience, reduced contextual stimuli for making decisions, and increased costs (e.g., purchasing and maintaining cameras, wireless microphones, playback devices, and monitors). Personnel time for data collection is more than doubled when using videos, because time is needed to analyze segments and set up and record them. Additionally, some participants and institutions may be reluctant to provide consent for videotaping, and participant reactivity is greater when being videotaped than when simply being observed.

Concluding Remarks

For more than three decades, I have used direct (systematic) observation to generate data to assess physical activity. This method: (a) provides contextual data on the activity setting, (b) is useful in examining how physical and social environments influence physical activity, and (C)
is assesses group physical activity in open environments, such as park and recreation settings. I believe systematic observation methods are underused by researchers; thus, this paper provides an overview of the methods and technology used as well as some practical examples from selected instruments.

**References**


**Note**

1. Protocols for the BEACHES, SOFIT, SOPLAY, and SOPARC observation systems may be downloaded from Active Living Research, San Diego (http://www.active-livingresearch.org/). Copies of the DVDs for training SOFIT and SOPLAY/SOPARC observers are also currently available free by sending a written request to Active Living Research.

**Author’s Notes**

I previously described concepts related to the systematic observation of physical activity in a chapter in *Physical Activity Assessments for Health-Related Research* (Welk, 2002). The current paper updates and extends that chapter. It is a tremendous honor to have been selected as the 2009 McCloy Lecturer. This achievement would not have been possible without the assistance, guidance, and support of many diverse, internationally recognized scholars. Among them are: Brent Rushall, who directed my master’s thesis at Dalhousie University and was the first to insist I put “my data where my mouths are;” Daryl Siedentop, director of my doctoral dissertation at The Ohio State University; Jim Sallis, professor of psychology at San Diego State University, Director of Active Living Research, and my health behavior mentor; Phil Nader, former head of Community Pediatrics at the University of California, San Diego; John Elder, Graduate School of Public Health at San Diego State University; and Deborah Cohen, RAND Corporation, Santa Monica, CA. I also wish to acknowledge Hans van der Mars and Phillip Ward, who provided comments on the original version of the manuscript. Please address correspondence concerning this article to Thomas L. McKenzie, San Diego State University, 5127 Walsh Way, San Diego, CA 92115-1148.

E-mail: tmckenzie@sdsu.edu