A Novel Video-Based Direct Observation System for Assessing Physical Activity and Sedentary Behavior in Children and Young Adults

Melanna F. Cox, Greg J. Petrucci Jr., and Robert T. Marcotte Brittany R. Masteller
University of Massachusetts Amherst Smith College

John Staudenmayer, Patty S. Freedson, and John R. Sirard
University of Massachusetts Amherst

Purpose: Develop a direct observation (DO) system to serve as a criterion measure for the calibration of models applied to free-living (FL) accelerometer data. **Methods:** Ten participants (19.4 ± 0.8 years) were video-recorded during four, one-hour FL sessions in different settings: 1) school, 2) home, 3) community, and 4) physical activity. For each setting, 10-minute clips from three randomly selected sessions were extracted and coded by one expert coder and up to 20 trained coders using the Observer XT software (Noldus, Wageningen, the Netherlands). The coder defines each whole-body movement which was further described with three modifiers: 1) locomotion, 2) activity type, and 3) MET value (used to categorize intensity level). Percent agreement was calculated for intra- and inter-rater reliability. For intra-rater reliability, the criterion coder coded all 12 clips twice, separated by at least one week between coding sessions. For inter-rater reliability, coded clips by trained coders were compared to the expert coder. Intraclass correlations (ICCs) were calculated to assess the agreement of intensity category for intra- and inter-rater comparisons described above. **Results:** For intra-rater reliability, mean percent agreement ranged from $91.9 \pm 3.9\%$ to $100.0 \pm 0.0\%$ across all variables in all settings. For inter-rater reliability, mean percent agreement ranged from $88.2 \pm 3.5\%$ to $100.0 \pm 0.0\%$ across all variables in all settings. ICCs for intensity category ranged from 0.74–1.00 and 0.81–1.00 for intra- and inter-rater comparisons, respectively. **Conclusion:** The DO system is reliable and feasible to serve as a criterion measure of FL physical activity in young adults to calibrate accelerometers, subsequently improving interpretation of surveillance and intervention research.

Keywords: accelerometers, activity classification, validation

A widely used tool to assess physical activity (PA) and sedentary behavior (SB) is the wearable accelerometer. Accelerometers are often used in free-living settings for surveillance and intervention studies. To quantify the amount and intensity of body movement, prediction models are applied to various features of the accelerometer data to estimate PA and SB.

Algorithms to estimate PA from accelerometer data often rely on laboratory calibration studies that use indirect calorimetry as a criterion measure for activity intensity. Laboratory calibration protocols require participants to complete structured or semi-structured activities in a controlled setting. However, algorithms developed from these laboratory calibration studies perform poorly when applied to different datasets or when applied to free-living data (Lyden, Keadle, Staudenmayer, & Freedson, 2014; Lyden, Kozey, Staudenmayer, & Freedson, 2011; Sasaki, Hickey, et al., 2016). Accelerometer calibration performed in free-living settings where devices will ultimately be used has the potential to improve prediction algorithms. However, an

appropriate and feasible free-living criterion measure must first be developed and tested.

Direct observation (DO) has the potential to be the free-living criterion measure. DO requires little equipment, is not restricted to the laboratory and has been identified as a criterion measure for categorizing free-living PA in children and adults (Kelly, Fitzsimons, & Baker, 2016; McKenzie, 2002; Sasaki, John, et al., 2016; Sirard & Pate, 2001). Further, unlike other criterion measures such as indirect calorimetry and doubly labeled water, DO provides information on the behavioral aspects of the movements (Fakhouri et al., 2014; Sasaki, John, et al., 2016). Several documented DO systems have been developed for use in children (Brown, Gooze, McIver, & Rathel, 2009; Brown et al., 2006; Cohen, McDonald, McIver, Pate, & Trost, 2014; McIver, Brown, Pfeiffer, Dowda, & Pate, 2009, 2016; McKenzie et al., 1991) and fewer in adults (Sasaki, John, et al., 2016; Welch, Swartz, Cho, & Strath, 2016) to assess the context, intensity and duration of PA. Therefore, DO has potential to be used as a criterion measure for free-living accelerometer calibration studies (McKenzie, 2002; Sirard & Pate, 2001).

Several DO systems exist that are designed to describe general PA in individuals and groups (McKenzie, 2002), but the use of these current DO systems as criterion measures for device calibration in free-living settings may not be appropriate. Although some DO systems could use continuous sampling, most of the preexisting DO systems use momentary sampling, characterized by repeated, fixed duration cycles of observing and recording.

Cox, Petrucci, Marcotte, Freedson, and Sirard are with the Department of Kinesiology; Staudenmayer is with the Department of Mathematics and Statistics; University of Massachusetts Amherst, Amherst, MA. Sirard is also with the Commonwealth Honors College, University of Massachusetts Amherst. Masteller is with the Department of Exercise and Sport Studies, Smith College, Northampton, MA. Cox (mfcx@umass.edu) is corresponding author.

An appropriate criterion measure for free-living calibration studies would need to provide data continuously to better match the short, instantaneous signals from the accelerometer. Therefore, momentary sampling is not ideal because as the researcher records the previously observed event, data are being missed and potentially misclassified. Alternatively, focal sampling is a continuous sampling process in which the recording of an event is initiated by each change in behavior (Welch et al., 2016). Therefore, researchers have begun to develop focal DO systems that can potentially be used as criterion measures for free-living calibration of devices.

A focal DO system was created to be used in adults (Sasaki, John, et al., 2016) and was assessed using indirect calorimetry in a simulated free-living setting (Lyden, Petruski, Staudenmayer, & Freedson, 2014). Lyden et al. reported no significant difference for SB, moderate, and vigorous PA intensity categories between indirect calorimetry and the DO system. Welch and colleagues extended this work into the free-living setting in older adults (Welch et al., 2016). Approximately 45% and 60% of metabolic equivalents (METs) estimated from the DO system were within 0.5 METs of the measured MET values for lab and free-living settings, respectively. The results of both studies suggest that the use of DO systems in the free-living settings could serve as a criterion measure for the calibration of models applied to accelerometer data. However, Lyden et al. and Welch et al. both incorporated laboratory settings which minimize the generalizability of their DO systems. Further, although Welch et al. did observe participants in the free-living setting, the participants were older adults which also limits the generalizability. Also, observations from both studies were completed in real-time. Real-time observation can compromise the accuracy of the observation due to observer fatigue and human error. To build off this previous research, 1) data should be collected in a free-living setting, 2) in a general adult population, and 3) use video-based observation to allow observers to pause, rewind, save and return to observations as needed.

Developing and testing a comprehensive, video-based, focal sampling DO system in free-living settings may provide a robust criterion measure for calibration of devices designed for free-living movement assessment. As a first step, it is important to gain an understanding of the performance of such a DO system used in free-living settings. The purpose of this study was to develop a novel, video-based, focal sampling DO system and examine the intra- and inter-rater agreement of trained coders.

criterion measure for the Movement Observation in Children and Adolescents (MOCA) Study. The MOCA Study will collect free-living accelerometer data and use DO as the criterion measure to ultimately develop valid accelerometer prediction models to estimate PA and SB. The current study was conducted in three phases (Figure 1). In phase I, an iterative development of the DO system was performed. During phase II, the initial agreement analyses were conducted with data collected from a pilot study in young children. This initial agreement analyses were completed in younger children because the intermittent nature of child free-play activity (Bailey et al., 1995) provided variability in movements and a challenging set of training videos for video coders. Based on these initial agreement analyses and identification of problematic issues, the DO system was refined (Figure 1, Phase II). In phase III, the final agreement analyses were conducted with data from the MOCA Study's young adult subsample (Figure 1, Phase III).

Phase 1: Iterative Development of the Direct Observation System

The DO system was developed and implemented using the Observer XT (OXT) (Noldus Information Technology, Inc., Wageningen, Netherlands) computer software to code video observations. Prior to any data collection, the DO system was developed using the previously mentioned adult DO systems as a template (Lyden, Petruski, et al., 2014). The first author (expert coder) took the lead on developing the DO system, with regular input provided by the research team.

The DO system uses focal sampling and includes codes for four main outcomes. For this focal sampling protocol, a new event is recorded each time a participant changes a component of their movement for at least one second. The main outcomes included whole-body movement, locomotion, activity type and the metabolic equivalent (MET) value for the whole-body movement. The whole-body movements include 12 options for different movements and postures. The final operational definitions are provided in Table 1. The other three outcomes (i.e., modifiers) include locomotion, activity type, and MET value. Locomotion is a binary "Yes" or "No" modifier that specifies if the participant moved from an initial position to another position with at least two steps. Activity type has several options for the context of the movement (e.g., running as whole-body movement with soccer as the activity type). The MET values in the pilot phase (child participants) were originally based on the MET values from the Compendium of Energy Expenditures in Youth corresponding with each whole-body movement (Ridley, Ainsworth, & Olds, 2008) and then updated using the most recent Youth Compendium of Physical Activities for children 6 to 17 years of age (Butte et al., 2018).

Methods

Study Design

This study was conducted to develop and assess the agreement between observers using the DO system that will be used as the

Figure 1 — Flowchart of study design.

Table 1 Operational Definitions and Guidelines for Whole-Body Movements

Whole-Body Movements	Operational Definition	Cue Start of Whole-Body Movement
Lying	Either the back, chest, or stomach is used as support; body is in horizontal or supine position	One of the areas mentioned above is fully on the surface
Sitting	Buttocks used as support on a surface	Initial contact between the buttocks and surface
Standing	Body is in an upright position; feet are used as support	Once feet are planted and body is in an upright position
Walking/Slow Walking	Moving from one point to another taking at least two steps	When heel of swing foot surpasses stance foot
WalkLoad	Walking with a load weighing least 2 lbs; or bag at least the size of a laptop	Heel of swing foot surpasses stance foot
Running	Locomotion with a flight phase; at one point, both feet are off the ground	Heel of swing foot surpasses stance foot
Kneeling	At least one knee is used as support on the surface; no upper limb support	At least one knee is fully on the surface
Skiping	Hopping off one foot and landing on the same foot	The body is in the lowest position right before take off
Climbing	Body is in a vertical position; hands and feet are used to travel vertically	At least one hand and one foot makes contact with the surface
Crawling	Torso is in a horizontal position; at least 3 bases of limb support through the duration of the movement	Three bases of support are fully on the surface
Squatting	Feet are used as support on the floor, thighs are parallel to the floor and buttocks are level with the knees or lower*	Immediately before buttocks begins to lower
Jumping	Feet, or one foot, used to propel upwards, and/or forward; there is a loading phase and in air phase	Behavior begins when the body is in the lowest position right before take-off**
Dancing	Feet or body, or both, move rhythmically in a pattern of steps	Initial rhythmic movement
Elliptical	Participant is using an Elliptical Trainer	Once movement begins with both feet on the elliptical trainer
Private Time***	Any time it is not appropriate to record the participant; i.e., using the restroom, changing clothes	Behavior begins when the participant goes into the private area/room
Obstructed View***	Any time the view of the participant is obstructed; i.e., behind playground apparatus	Behavior begins when the participant is no longer in view

*Special case—at the gym, people may not be properly doing a squat but the behavior should still be coded as squatting. **The loading phase (will look like a squat but is not coded as this will usually not last for more than one second. ***These “whole body movements” will not require you to code any modifiers.

The MET value modifier for the young adult sub-sample (Phase III) was based on the Compendium of Physical Activities (Ainsworth et al., 2011). The options within each modifier were mutually exclusive and required for every new event.

Phase 2: Pilot Study

All methods of the pilot study were first approved by the University of Massachusetts Amherst Institutional Review Board. A parent/guardian provided informed consent and the child provided their written assent to participate in the study. The pilot study was conducted in 28 children (8.4 ± 1.5 years, 28% female). Each participant completed a 30-minute indoor free-play session that was recorded by a research assistant using a GoPro camera (GoPro, San Mateo, CA). The participants were informed that they could participate in any activities during the duration of the session. There was a range of activities and toys provided including an obstacle course, balls, blocks, and puzzles. A research assistant followed the child with the camera for the duration of the session. The videos were then downloaded and opened in the OXT for coding. All video-recorded free-play sessions were coded for the whole-body movement and modifiers in the OXT software by the expert coder ($N = 28$ videos). Six trained coders coded a subset of the videos and were used to assess inter-rater agreement. The subset of videos included six, randomly selected 5-minute clips. Intra- and inter-rater

agreement was calculated using second-by-second agreement of each whole-body movement and all modifiers. Data processing methods used in the pilot study were the same methods used in the main study (see Phase III, Data Processing). Briefly, intra-rater agreement was acceptable ($>80\%$) for all variables except MET value ($69 \pm 27\%$). Inter-rater agreement ($n = 6$ coders compared with expert coder) ranged from 50%–95% agreement for whole-body movement, locomotion, activity type, and METs.

Direct Observation Refinement

Based on the results from the pilot study, changes were made to coder training and the rules for the DO system were further clarified to arrive at the final DO system. Whole-body movement was widely affected by interpretation of ambiguous body positions. For example, a child could be standing but have one knee on a chair. Based on previous definitions, coders could interpret the whole-body movement as either standing or kneeling. A guideline was then implemented which required standing to be coded when at least one foot was on floor regardless of leaning on other limbs (e.g., knees, hands, elbows).

The default use of the whole-body movement MET values from the Compendium of Energy Expenditures for Youth was problematic (Butte et al., 2018; Ridley et al., 2008). For example, kicking a soccer ball while stationary was coded as 1.5 MET

because the whole-body movement was coded as standing. Since the participant was observed doing a moderate intensity activity (kicking a soccer ball), using the MET values solely on the whole-body movement did not appropriately reflect the intensity of the activity. Thus, a new guideline was implemented that required coders to decide whether to use the MET value associated with the whole-body movement or the activity type (e.g., soccer MET value) based on observed intensity.

The activity type modifier in the DO system included most activities in the original Compendium of Energy Expenditures for Youth (Ridley et al., 2008). Coders had several choices of activities for each second of the observation, which increased variability in interpretation leading to a likely source of low agreement. Several activities listed in the Compendium were appropriate for the same situation. For example, if someone is typing and reading a document, the behavior could be coded as both “Computer Work” and “Reading” to accurately describe the activity type. We chose not to shorten the list of activities in efforts to preserve the robustness of the DO system. Instead, to refine our DO system, distinct criteria were developed across similar activity type modifiers in the Compendium. For example, computer work was coded as the activity type if a participant’s hands were on the computer despite them reading at the same time. The rationale for this was because all of the movement was occurring with computer work.

Inter-rater agreement for the MET value modifier varied widely across videos (39%–67%). However, upon further inspection of the videos, there were often discrepancies in coded observations when participants were observed shifting their weight while standing and typically involved only 1–2 consecutive steps. Consequently, the coders either recorded a 1.5 MET for “Standing” or a “2.9” for light “Walking”. Therefore, a guideline was implemented that stated a participant must take two or more purposeful steps for the behavior to be considered “Walking” versus “Standing”. Final operational definitions and guidelines for whole-body movements are presented in Table 1. Locomotion was defined as a participant moving from one place to another taking at least 2 steps.

Phase 3: Main Study

Participants. Recruitment for the main MOCA study took place in the Western Massachusetts area. Participants included 18- to 24-year-old individuals free from any illnesses or disabilities that could limit their ability to participate in physical activity. All participants read and provided a written informed consent form approved by the Institutional Review Board of the University of Massachusetts Amherst.

Procedures. Participants completed four, one-hour observation sessions. Each session was in a distinct free-living setting, which included home, school, community, and physical activity/exercise. Anthropometrics and demographics were collected at the initial session. Height and weight were measured to the nearest tenth of a centimeter (cm) and tenth of a kilogram (kg) using a portable stadiometer (Shorr Board, Weigh and Measure, LLC, Olney MD) and digital scale (SECA 877), respectively. Participants self-reported age, sex and race/ethnicity. Each session was video-recorded using a GoPro HERO camera (GoPro, San Mateo, California). The clock from the screen of a study laptop was recorded on video to display each session’s start time. Once the session start time and participant were clearly in view in the video recording frame, the participant was free to perform any activities for the remainder of the session. At the end of each session, the participant and laptop clock were again placed in view of the video recording frame to mark the end of the

session. All videos used in the agreement analyses were coded in the OXT software. The refined DO system rules and definitions created in Phase II were used to code all videos (Table 1).

Coder Training

Each coder ($n = 20$) completed ~30 hours of formal training with the principle investigator, expert coder, and other lab personnel. Training included review and discussion of literature relevant to the DO of physical activity in children, familiarization to the OXT software, and coding of 6 training videos. Training videos were randomly selected 5-minute clips from the pilot study. The pilot study videos (young children in a free-play setting) were used for training because the intermittent movement patterns (Bailey et al., 1995) facilitated a more challenging training protocol. By starting with more difficult videos and a rigorous training protocol, the coders were better prepared to demonstrate satisfactory inter-rater agreement (>80%) when observing the less sporadic movements seen in the young adult sample.

Coder Certification

After training, the coders were certified prior to coding videos from the MOCA study. Videos from the pool of young adult data collection sessions ($n = 50$ videos) were stratified by environment and three videos from each environment were randomly selected and a randomly chosen 10-minute clip was trimmed from each hour-long video. To become certified in each of the four free-living environments, coders coded the three, 10-minute video clips (certification videos) obtained during each of the four settings (12 total video clips). Research assistants coded the first activity at the start of the video to ensure the activity coded aligned with the video. The observations from the coders were compared to the corresponding observations of the expert coder. Beginning with the school setting, coders had to obtain a percent agreement of 80% or higher when compared to the expert coder across all direct observation variables: 1) whole-body movement, 2) locomotion, 3) activity type, 4) MET value, and 5) intensity category. Once a coder obtained a percent agreement of 80% or higher across all variables for all three videos in the school setting, the process was repeated for home, community, and physical activity/exercise settings. The progression through settings was selected based on the increasing difficulty of the videos. Failure to obtain a percent agreement of at least 80% resulted in further training based on identification of coding errors within the certification videos and then subsequent attempts at certification. Only three attempts were allotted for each setting; after a third failure the coder was ineligible for certification for that setting. All coders from the current study were certified with three or fewer attempts.

Data Processing

The same data processing methods were used in the pilot study and main study. Specifically, the output from each observation was exported from Noldus OXT to a comma delimited (.CSV) file and imported into RStudio (R: A language and environment for statistical computing, R Foundation for Statistical Computing, 2017, Boston, MA) for additional data processing. Files from the OXT produced only a single data point for an event which does not allow for second-by-second analyses. Therefore, each .CSV file was modified in R by generating a new file that created a data point for each second of the observation, instead of a data point for each event. For example, the original data file may indicate five

seconds spent sitting as one data point. The modified file would read one second spent sitting for five consecutive rows in the data file. This second-by-second modified DO file was used for all analyses. PA intensity categories were calculated using the MET value for each 1-second epoch. One-second MET values were categorized as SED (sedentary, 1.0–1.49 METs), LPA (light physical activity, 1.5–2.9 METs) MPA (moderate physical activity, 3.0–5.9 METs), or VPA (vigorous physical activity, 6.0 ≥ METs).

Statistical Analyses

The certification video clips (*n* = 12) were used to perform agreement analyses. The expert coder re-coded all 12 videos (three certification videos from all four settings) at least seven days apart to assess intra-rater agreement. The trained coders (*n* = 20) coded the same twelve certification videos to assess inter-rater agreement. Percent agreement is used to assess intra- and inter-rater agreement to produce comparable results to previous studies. Intra-rater (expert coder only) and inter-rater agreement (trained coders compared to expert coder) were assessed using second-by-second percent agreement for all variables: whole-body movement, locomotion, activity type, MET value, and intensity category. Percent agreement between the expert coder’s initial observation and second observation of each certification video was used to assess intra-rater agreement. Percent agreement between each observation from trained coders and the expert coder’s corresponding observation were used to assess inter-rater agreement (Figure 2). Intraclass correlations were calculated for intensity category to assess intra-rater and inter-rater agreement. All analyses were completed in Rstudio.

Results

Descriptive Statistics

The 12 videos randomly selected for the certification videos represented 10 unique participants. Participants (*n* = 10) were 80% female, 19.4 ± 0.8 years old, 166.4 ± 4.4 (cm) and weighed 61.6 ± 7.3 (kg).

Intra-Rater Agreement

Percent agreement ranged from 91.9% ± 3.9% to 100% ± 0 across all variables in all environments (Table 2). Variables derived from the home environment videos had the overall highest percent agreements (93% to 100%) and those from the community environment videos resulted in the lowest percent agreements (91.9% to 96.4%). Intraclass correlation coefficients for intensity category were 1.00 ± 0.00, 1.00 ± 0.00, 0.86 ± 0.09, 0.96 ± 0.00 for school, home, community, and physical activity settings, respectively.

Inter-Rater Agreement

A total of 20 coders completed training and were certified for school settings, 15 were certified for home settings, eight were certified for community settings, and five were certified for the physical activity/exercise setting. Percent agreement ranged from 88.2% ± 3.5% to 100% ± 0% across all variables and environments (Table 3). Overall, variables derived from the school environment had the highest inter-rater percent agreement and those from the community environment had the lowest percent agreement across

Figure 2 — Visual example of data used to calculate inter-rater agreement.

Table 2 Intra-rater Percent Agreement for All Variables by Environment ($M \pm SD$)

Setting	Whole-Body Movement	Locomotion	Activity Type	METs	Intensity Category
School	100% $\pm$ 0%	100% $\pm$ 0%	96.9% $\pm$ 5.1%	99.6% $\pm$ 0.7%	100% $\pm$ 0%
Home	100% $\pm$ 0%	100% $\pm$ 0%	93.5% $\pm$ 5.0%	99.6% $\pm$ 0.4%	100% $\pm$ 0%
Community	95.4% $\pm$ 2.2%	96.4% $\pm$ 1.8%	94.2% $\pm$ 8.1%	91.9% $\pm$ 3.9%	94.5% $\pm$ 4.2%
Physical Activity	98.8% $\pm$ 1.1%	99.6% $\pm$ 0.2%	99.7% $\pm$ 0.3%	98.1% $\pm$ 2.3%	98.6% $\pm$ 1.7%

Table 3 Inter-rater Percent Agreement for All Variables by Environment ($M \pm SD$)

Setting	Whole-Body Movement	Locomotion	Activity Type	METs	Intensity Category
School	99.8% $\pm$ 1.0%	100% $\pm$ 0.1%	94.8% $\pm$ 5.8%	99.9% $\pm$ 0.3%	100% $\pm$ 0%
Home	100% $\pm$ 0.1%	100% $\pm$ 0%	88.8% $\pm$ 5.0%	99.5% $\pm$ 1.0%	99.8% $\pm$ 0.9%
Community	92% $\pm$ 4.0%	94.7% $\pm$ 3.4%	92.6% $\pm$ 6.8%	88.2% $\pm$ 3.5%	91.6% $\pm$ 4.1%
Physical Activity	93.2% $\pm$ 6.1%	98.1% $\pm$ 2.2%	93.7% $\pm$ 5.4%	91.0% $\pm$ 6.5%	92.9% $\pm$ 5.8%

all variables. Intraclass correlation coefficients for intensity category were 1.00 ± 0.00 , 1.00 ± 0.00 , 0.81 ± 0.09 , and 0.85 ± 0.16 for school, home, community, and physical activity/exercise settings, respectively.

Discussion

The purpose of this study was to develop and evaluate a novel, video-based, focal sampling DO system for use as a criterion measure for accelerometer calibration studies. The system was iteratively developed and rigorously assessed for agreement among observers and feasibility of implementation on free-living data. The design of the system and use of the OXT software provides a feasible tool to capture instantaneous movement and the context of those movements. Overall, intra- and inter-rater agreement were high across all variables indicating a feasible and reliable DO system.

The current video-based, focal sampling DO system was expected to show higher intra- and inter-rater agreement than past DO systems relying on real-time momentary sampling. Intra-rater agreement was similar between past DO systems ($ICC = 0.70$ – 0.86) (McKenzie, Marshall, Sallis, & Conway, 2000) and the MOCA DO system ($ICC = 0.74$ – 1.00). Inter-rater agreement was above 80% for most earlier studies and the MOCA DO system (91.6%–100%). These results suggest that both the previous DO systems and the MOCA DO system are similar in agreement. When DO is completed in real-time, less complexity and detail are possible. However, the increase in complexity and detail available with the video-based MOCA DO system may also attenuate agreement statistics. The comparison of previous momentary sampling DO systems to the MOCA DO system is not ideal due to the different purposes of these systems. However, these comparisons provide a basis for the agreement of the MOCA DO system since there is only one other study that developed a focal DO system designed for device calibration in adults.

As mentioned above, the only other physical activity DO system developed using an adult population, and the most comparable to the DO system presented in this paper, is that from Sasaki, John, et al., 2016. Although Sasaki, John, et al., (2016) do not report intra-rater reliability, they do report inter-rater reliability for both activity intensity category and a whole-body movement variable analogous to the one reported for the current system.

Inter-rater agreement from the current study for intensity category ($91.6\% \pm 4.1\%$ to $100\% \pm 0.0\%$) and whole-body movement ($92.0\% \pm 4.0\%$ to $100\% \pm 0.1\%$) were stronger and less variable than those from the DO system created by Sasaki and colleagues (intensity category $76.1\% \pm 15.4\%$; and whole-body movement $86.6\% \pm 6.5\%$). The DO system developed by Sasaki et al. was used during real-time observations. In contrast, the current DO system uses recorded videos and the OXT software allowing the coders to rewind, slow down, or pause the observations. Therefore, the improved inter-observer agreement with the current DO system can likely be attributed to not having to code an individual's behavior in real-time and the coders' ability to check their coding before submitting for agreement analyses.

Intra-rater reliability for intensity category from the current DO system ranged from 0.74 to 1.00 across all 12 videos coded by the expert coder. Intra-rater agreement has been reported for only one other DO system; the System for Observing Play and Leisure Activities (SOPLAY) (McKenzie et al., 2000). The SOPLAY system has been widely used and was tested by an independent research group (Saint-Maurice, Welk, Ihmels, & Krapfl, 2011). Intra-class correlations for intensity category ranged from 0.70 to 0.86, based on three-day test-retest reliability in three observers coding training videos (Saint-Maurice et al., 2011). The generally stronger intra-rater agreement for the current DO system, compared to those from the SOPLAY system, is likely due to the video-based nature of the current system. The increased complexity of the current DO system could lead to lower agreement scores since intensity category is derived from a recorded MET value as opposed to a three-category option in the SOPLAY system. However, the complexity of the current system is offset by using video recordings and the OBX software, which allows for careful examination of each movement; a level of precision not possible when coding someone's movement behaviors in real-time.

Strengths

The current study has several strengths including an iterative development process, rigorous agreement analysis and a wide range of participant ages for DO system development. The MOCA DO system was designed to be a criterion measure for free-living calibration of accelerometers used to estimate PA and SB across a wide age range. Therefore, the development process to create the DO system was meticulous and thorough to provide

a high-quality criterion measure. Also, the method of DO from the MOCA system has the potential to be applied to all populations, across a wide range of settings and activities.

The agreement analysis in the current study is unique due to the large number of coders being compared ($N=20$ for inter-rater agreement) and the second-by-second agreement analyses. The strong inter-rater agreement across multiple coders is a result of an extensive training process that can be used with a variety of people from diverse backgrounds and skillsets. The second-by-second analysis for intra- and inter-rater agreement also demonstrates the robust nature of the current DO system. Accelerometers can measure movement at very fine time intervals. Since the DO system will be used to develop algorithms based on the accelerometer data, it was imperative that the output of the DO system produce precise measures of movement.

Lastly, a strength of this study is that indirect calorimetry was not used as the criterion measure for movement. It is well-known that the performance of models, calibrated using indirect calorimetry in structured and semi-structured settings, are attenuated when they are used in free-living settings (Lyden, Keadle, et al., 2014; Lyden et al., 2011; Sasaki, Hickey, et al., 2016). The sporadic, intermittent nature of free-living data creates a physiological response that lags behind the movement behavior. Accelerometers capture instantaneous movement and, therefore, should be calibrated using a criterion measure for movement, such as the DO system introduced in this paper. It is important to note that the current DO system uses MET values from the compendium to estimate intensity categories, not to calculate point estimates of energy expenditure. Therefore, the MOCA DO system should not be used as a criterion measure for energy expenditure.

Limitations

Despite the strengths of this study, there are some limitations. The pilot study was conducted in 6- to 10-year-old children while the main study was conducted in young adults. While this may seem inconsistent, the videos from the pilot study were intentionally difficult and were completed before additional guidelines were adopted for coding videos from young adult participants in the MOCA Study. Moving forward, coders will be required to recertify for different age groups, such as toddler (1.5–2.9 years), preschool (3–5.9 years), and 6- to 12-year-old age groups, before coding videos from those participants. Therefore, coders will be required to complete certification videos for younger age groups before coding data from the main study, requiring additional training. For these age groups, the new compendium will provide somewhat broad age-based adjustments to the coded MET value. Although broad, this age-based adjustment is appropriate for our goal of estimating the intensity category. However, it should be noted that some measure of maturation would be needed for more accurate point estimates of EE. Another limitation to this study is the smaller number of coders certified across all domains. Fewer coders certified for the community and activity environments was primarily a factor of time available during a university semester to complete training and certifications, while also contributing to data collection efforts. Fewer coders may result in higher percent agreements than those that would be observed with more coders. Lastly, a drawback of collecting data during four one-hour free-living sessions is that we may not have captured rare activities or other unique ways the individual may be using their time. However, we collected data in four different activity domains, which represented a wide variety of activity types and intensities used for

the current study. These data will also provide a robust data source for future device calibration work.

Conclusion

Although accelerometers are frequently used in a free-living setting, there is currently no well-documented criterion measure for free-living calibration of these devices. The MOCA DO coding system is a feasible and reliable criterion measure for free-living movement in young adults. Therefore, the DO system could be used as a standardized criterion measure to calibrate a variety of accelerometer devices using free-living data. Thus, the DO system should have positive impacts on accelerometer classification precision and accuracy in the free-living setting and will subsequently improve how we assess the habitual PA of different populations.

Acknowledgments

A grant from the National Institutes of Health supported this study (NIH NIDDK 1R01DK110148). The authors would like to thank the participants for their time and the research assistants for their dedication to providing quality direct observation data.

References

- Ainsworth, B.E., Haskell, W.L., Herrmann, S.D., Meckes, N., Bassett, D.R., Jr., Tudor-Locke, C., . . . Leon, A.S. (2011). 2011 compendium of physical activities: A second update of codes and MET values. *Medicine & Science in Sports & Exercise*, 43(8), 1575–1581. doi:10.1249/MSS.0b013e31821ecce12
- Bailey, R.C., Olson, J., Pepper, S.L., Porszasz, J., Barstow, T.J., & Cooper, D.M. (1995). The level and tempo of children's physical activities: An observational study. *Medicine & Science in Sports & Exercise*, 27(7), 1033–1041. doi:10.1249/00005768-199507000-00012
- Brown, W.H., Googe, H.S., McIver, K.L., & Rathel, J.M. (2009). Effects of teacher-encouraged physical activity on preschool playgrounds. *Journal of Early Intervention*, 31(2), 126–145. doi:10.1177/1053815109331858
- Brown, W.H., Pfeiffer, K.A., McIver, K.L., Dowda, M., Almeida, M.J., & Pate, R.R. (2006). Assessing preschool children's physical activity: The observational system for recording physical activity in children-preschool version. *Research Quarterly for Exercise and Sports*, 77(2), 167–176. doi:10.1080/02701367.2006.10599351
- Butte, N.F., Watson, K.B., Ridley, K., Zakeri, I.F., McMurray, R.G., Pfeiffer, K.A., . . . Fulton, J.E. (2018). A youth compendium of physical activities: Activity codes and metabolic intensities. *Medicine & Science in Sports & Exercise*, 50(2), 246–256. doi:10.1249/MSS.0000000000001430
- Cohen, A., McDonald, S., McIver, K., Pate, R., & Trost, S. (2014). Assessing physical activity during youth sport: The observational system for recording activity in children: Youth sports. *Pediatric Exercise Science*, 26(2), 203–209. PubMed ID: 24277926 doi:10.1123/pes.2013-0095
- Fakhouri, T.H., Hughes, J.P., Burt, V.L., Song, M., Fulton, J.E., & Ogden, C.L. (2014). Physical activity in U.S. youth aged 12–15 years, 2012. *NCHS Data Brief*, (141), 1–8. PubMed ID: 24401547
- Kelly, P., Fitzsimons, C., & Baker, G. (2016). Should we reframe how we think about physical activity and sedentary behaviour measurement? Validity and reliability reconsidered. *The International Journal of Behavior Nutrition and Physical Activity*, 13, 32. doi:10.1186/s12966-016-0351-4

- Lyden, K., Keadle, S.K., Staudenmayer, J., & Freedson, P.S. (2014). A method to estimate free-living active and sedentary behavior from an accelerometer. *Medicine & Science in Sports & Exercise*, 46(2), 386–397. doi:[10.1249/MSS.0b013e3182a42a2d](https://doi.org/10.1249/MSS.0b013e3182a42a2d)
- Lyden, K., Kozey, S.L., Staudenmayer, J.W., & Freedson, P.S. (2011). A comprehensive evaluation of commonly used accelerometer energy expenditure and MET prediction equations. *European Journal of Applied Physiology*, 111(2), 187–201. doi:[10.1007/s00421-010-1639-8](https://doi.org/10.1007/s00421-010-1639-8)
- Lyden, K., Petruski, N., Staudenmayer, J., & Freedson, P. (2014). Direct observation is a valid criterion for estimating physical activity and sedentary behavior. *Journal of Physical Activity and Health*, 11(4), 860–863. PubMed ID: [25078528](https://pubmed.ncbi.nlm.nih.gov/25078528/) doi:[10.1123/jpah.2012-0290](https://doi.org/10.1123/jpah.2012-0290)
- McIver, K.L., Brown, W.H., Pfeiffer, K.A., Dowda, M., & Pate, R.R. (2009). Assessing children's physical activity in their homes: The observational system for recording physical activity in children-home. *Journal of Applied Behavior Analysis*, 42(1), 1–16. PubMed ID: [19721726](https://pubmed.ncbi.nlm.nih.gov/19721726/) doi:[10.1901/jaba.2009.42-1](https://doi.org/10.1901/jaba.2009.42-1)
- McIver, K.L., Brown, W.H., Pfeiffer, K.A., Dowda, M., & Pate, R.R. (2016). Development and testing of the observational system for recording physical activity in children: Elementary school. *Research Quarterly for Exercise and Sports*, 87(1), 101–109. doi:[10.1080/02701367.2015.1125994](https://doi.org/10.1080/02701367.2015.1125994)
- McKenzie, T., Sallis, J.F., Patterson, T., Patterson, T.L., Elder, J.P., Berry, C.C., . . . Nelson, J.A. (1991). BEACHES: An observational system for assessing children's eating and physical activity behaviors and associated events. *Journal of Applied Behavior Analysis*, 24(1), 141–151. doi:[10.1901/jaba.1991.24-141](https://doi.org/10.1901/jaba.1991.24-141)
- McKenzie, T.L. (2002). The use of direct observation to assess physical activity. In: Welk, G., ed. *Physical activity assessments for health-related research* (pp. 179–195). Champaign, IL: Human Kinetics.
- McKenzie, T.L., Marshall, S.J., Sallis, J.F., & Conway, T.L. (2000). Leisure-time physical activity in school environments: An observational study using SOPLAY. *Preventive Medicine*, 30, 70–77. PubMed ID: [10642462](https://pubmed.ncbi.nlm.nih.gov/10642462/) doi:[10.1006/pmed.1999.0591](https://doi.org/10.1006/pmed.1999.0591)
- Ridley, K., Ainsworth, B.E., & Olds, T.S. (2008). Development of a compendium of energy expenditures for youth. *The International Journal of Behavioral Nutrition and Physical Activity*, 5(1), 45. doi:[10.1186/1479-5868-5-45](https://doi.org/10.1186/1479-5868-5-45)
- Saint-Maurice, P.F., Welk, G., Ihmels, M.A., & Krapfl, J.R. (2011). Validation of the SOPLAY direct observation tool with an accelerometer-based physical activity monitor. *Journal of Physical Activity and Health*, 8(8), 1108–1116. PubMed ID: [22039129](https://pubmed.ncbi.nlm.nih.gov/22039129/) doi:[10.1123/jpah.8.8.1108](https://doi.org/10.1123/jpah.8.8.1108)
- Sasaki, J.E., Hickey, A.M., Staudenmayer, J.W., John, D., Kent, J.A., & Freedson, P.S. (2016). Performance of activity classification algorithms in free-living older adults. *Medicine & Science in Sports & Exercise*, 48(5), 941–950. doi:[10.1249/MSS.0000000000000844](https://doi.org/10.1249/MSS.0000000000000844)
- Sasaki, J.E., John, D., Hickey, A.M., Lyden, K., Hagobian, T., & Freedson, P.S. (2016). Feasibility of using a continuous direct observation technique for assessment of free-living physical activity in young adults. *Archives of Sports Sciences*, 4(1), 2–6.
- Sirard, J.R., & Pate, R.R. (2001). Physical activity assessment in children and adolescents. *Sports Medicine*, 31(6), 439–454. PubMed ID: [11394563](https://pubmed.ncbi.nlm.nih.gov/11394563/) doi:[10.2165/00007256-200131060-00004](https://doi.org/10.2165/00007256-200131060-00004)
- Welch, W.A., Swartz, A.M., Cho, C.C., & Strath, S.J. (2016). Accuracy of direct observation to assess physical activity in older adults. *Journal of Aging and Physical Activity*, 24, 583–590. PubMed ID: [26964757](https://pubmed.ncbi.nlm.nih.gov/26964757/) doi:[10.1123/japa.2015-0216](https://doi.org/10.1123/japa.2015-0216)