Psychometric Properties of Two Systematic Observation Techniques for Assessing Physical Activity Levels in Children with Mental Retardation

Christina A. Taylor and Joonkoo Yun

This study examined the psychometric properties of the System for Observing Fitness Instruction Time (SOFIT) and the Children’s Activity Rating Scale (CARS) for use with children with mental retardation (MR). Eleven children with MR were videotaped while participating in a university-based community outreach program. Actiwatch accelerometers were used as the criterion measure. Results indicated that SOFIT and CARS both demonstrated adequate levels of generalizability ($\phi = 0.98$ and $0.75$), but a low concurrent validity coefficient for SOFIT ($r = .10$) and a moderate level of validity coefficient for CARS ($r = .61$) were observed. CARS demonstrates stronger validity evidence than SOFIT, but it is important to have sufficient rater training before using CARS for measuring physical activity level of children with MR.

Accurate information about physical activity levels obtained from a field-based method is important for health-related research, supervision, and curriculum development (8). Methods of measuring physical activity levels in field-based assessment not only should demonstrate good psychometric properties but also have a low cost and be easy to use in practical situations. Systematic observation is a field-based assessment method that is used to record behavior based on specifically designed guidelines and coding procedures (19). Information obtained through systematic observation may include the type, intensity, duration, physical environment, and social environment of activity (13). Many systematic observation tools, such as Children’s Physical Activity Form (13), Movement of Arms and Legs (7), and Activity Patterns and Energy Expenditure (2), have been developed to measure levels of physical activity in clinical settings for children without disabilities (1). However, McKenzie (8) stated “insufficient information exists to indicate which systems are most accurate and which can be used most reliably. . . . Simultaneous coding of children’s physical activity from videotapes using different observation systems is needed” (p. 227).

The System for Observing Fitness Instruction Time (SOFIT; 9) is an interval and momentary time sampling instrument, while the Children’s Activity Rating
Scale (CARS; 14) is a partial interval coding instrument. Both instruments provide methods for calculating a numerical summary score, which could be used for data analysis and evaluation of overall level of physical activity. These instruments have demonstrated good psychometric properties for children without disabilities and are commonly employed to measure children’s level of physical activity in both research and educational settings. However, there is limited information regarding the appropriateness of using these instruments for children with mental retardation (MR).

Psychometric properties of SOFIT for individuals with MR have been examined in a study by Hodge and Porretta (5), who report validity coefficient ranges from \( r = 0.06 \) to 0.90 between heart rates and SOFIT scores for eight children with MR. There are, however, a few methodological issues with Hodge and Porretta’s design. First, the heart rate monitor intervals and SOFIT intervals were not synchronized. The participant’s heart rate was measured every 15 s, but data from SOFIT were collected every 20 s. Second, no reliability information was cited. Third, the study did not use a conventional correlation method. Heart rate values and SOFIT levels for each person were correlated rather than group data; therefore, the reported validity coefficients were for each individual. These coefficients are a descriptive measure and may not be generalized. Hodge and Porretta argue that SOFIT has face validity for individuals with MR because there are no physical disabilities that would affect posturing (sitting, lying down) and subsequently scoring of this instrument. Considering individuals with MR demonstrate considerably less activity than individuals without MR (1,3,4), it is not known whether systematic observation will be able to discriminate among activity levels for this population. Yun and Ulrich (20) suggest that validity generalization should be limited to populations with similar behavioral characteristics.

Within our current knowledge, no study has examined psychometric properties for CARS with individuals with MR. Considering that accurately evaluating one’s level of physical activity in an educational or clinical setting becomes an important first step in planning and developing physical intervention programs for children with MR, it is important to identify appropriate field-based assessment tools that can provide researchers and practitioners a way to measure physical activity levels. Therefore, the purpose of this study was to evaluate the psychometric properties of systematic observation tools with individuals with MR to examine if reliability and validity evidence can be generalized to this population.

**Methods**

**Participants**

Participants in this study were 11 children between the ages of 6 and 14 years (\( M = 10.5, SD = 2.50 \) years) with MR, who participated in a university-based community outreach program. All participants of this program were diagnosed by a physician or school official, and parents indicated the participants’ disability on the program application. Participation was limited to ambulatory individuals because the selected systematic observation tools do not include coding procedures for wheelchair users. All participants had moderate to mild MR, though IQ scores were not obtained for the study. The study was approved by the investigators’
institutional review board. Written parental consent and participant verbal consent were obtained before data collection.

**Instruments and Apparatus**

Two systematic observation tools, CARS and SOFIT, and Actiwatch accelerometers, criterion measures for concurrent validity, were used to measure physical activity levels of participants. CARS was developed to categorize physical activities of various movement intensities. Intensities of movement range from no movement (Level 1) to very fast translocation (Level 5). An activity rating score was calculated for each minute by taking the mean of the activity levels recorded in the interval. Puhl et al. (14) reported interrater agreement of 84.1% for CARS. Validity evidence for CARS levels was investigated using VO$_2$ uptake and coding levels. Significant differences ($p < .05$) in the VO$_2$ uptake values for the five CARS levels demonstrate activity levels may be differentiated using the CARS system.

SOFIT is a momentary time sampling and interval recording systematic observation tool that divides physical activity behaviors into five categories ranging from lying down to very active. Levels 4 and 5 are equivalent to moderate-to-vigorous physical activity (MVPA; 15). An activity score can be computed using the percentage of the intervals in which the individual was performing MVPA. Physical activity behaviors are coded according to what the participant is doing at the moment the interval ends. SOFIT has demonstrated psychometric properties for individuals without disabilities. Rowe et al. (15) reported reliability across tasks ranging from $R = 0.80$ for jogging to $R = 0.91$ for curl-ups. Validity evidence was examined using heart rates across the different SOFIT levels (15). There was a significant task effect ($p < .01$).

Actiwatch (10) accelerometers were used as the concurrent measure for physical activity levels for comparison to CARS and SOFIT data. Accelerometers provide level of physical activity in the form of activity counts. The activity count represents the summed magnitude of the accelerations. This gives an indication of how much the individual moved in the interval recorded. Nichols, Morgan, Sarkin, Sallis, and Calfas (12) reported that 90% of the variance in energy expenditure was explained by the accelerometer activity counts.

**Procedures**

Data collection involved videotaping physical activity of participants during a gym-based program. The program follows a four-part lesson plan of an introductory activity, a fitness activity, a lesson focus, and a game. Videotaping took place during a variety of lesson contexts, such as hockey, lummi sticks, parachute, fitness activities, rhythmic gymnastics, and disc throwing. Each participant was videotaped for 15 min.

An accelerometer was placed on the nondominant hand of each participant upon arrival at the gym. Although the placement of an accelerometer on the hip may be more common, hand placement was selected because hip placement tends to underestimate energy costs of physical activities (18) and it is difficult to standardize a location of accelerometer on the hip in a real activity setting. Accelerometers were set to collect a movement count every 15 s. Hand dominance was determined one
of four ways: asking each child which hand he or she writes with, giving the child a pen and asking him or her to write his or her name on a piece of paper, asking the child’s parent, or asking the child’s clinician.

Coding of the videotapes using SOFIT and CARS was performed by three research assistants. Raters were provided with approximately 7 hr of training for the two systematic observation tools (CARS and SOFIT). Raters also practiced coding independently with multiple practice tapes before actual data collection. Training on each system followed the protocol outlined by van der Mars (19). Before actual data coding, intrarater reliability among raters was established. For CARS, intrarater reliability was calculated using the percentage of agreement for each interval among raters. Mean percentage of agreement for all intervals was then used as the estimate of intrarater reliability. Mean percentage of agreement between raters after training ranged from 87% to 93%.

Intrarater reliability for SOFIT was estimated using a strict interval-by-interval agreement procedure (11). The number of intervals agreed on was divided by the total number of intervals to obtain an estimate of intrarater reliability. After training, the intrarater reliability among the three raters was 0.90. Also, each rater’s performance was compared with an expert who had used SOFIT for 9 years and had conducted validity studies for the instrument. The intrarater reliability between the expert and the raters before data analysis ranged from 0.83 to 0.87.

Once coding actual data began, each tape was coded four times, twice each with SOFIT and CARS by each rater. SOFIT was coded using 15 s intervals to be synchronized with accelerometer data. CARS was coded using 1 min intervals and was also synchronized with 1 min summed accelerometer counts. A minimum 1-week delay was observed between coding. The coding of tapes was counterbalanced to reduce observer drift. Observer drift in systematic observation occurs when individuals change coding rules and interpret categories differently than the rules state (19). Observer drift was a concern in this study because raters switched between instruments frequently. The counterbalancing was set up as follows: (1) Rater A always coded SOFIT and then CARS; (2) rater B always coded CARS and then SOFIT; and (3) rater C coded SOFIT and then CARS the first time a participant was coded and coded with CARS and then SOFIT the second time a participant was coded. Counterbalancing ensures that SOFIT was coded before CARS as many times as CARS was coded before SOFIT.

Data Analysis

Reliability is estimated by using generalizability theory. Generalizability theory (g-theory) is a data analysis technique, which allows researchers to examine sources of variability in data. Traditional reliability analysis provides only a single score to evaluate the reliability of measurement. G-theory not only provides a single coefficient but also provides magnitudes of error from multiple sources (17). G-theory focuses on the magnitude of variability related to the participants, possible sources of error (facets), and their interactions. When researchers are able to pinpoint the sources of error, it provides them with information needed to improve the reliability of the measurement.

In this study, a two-facet generalizability study design was used to examine the sources of variability in assessing physical activity levels. For SOFIT, the percentage
of time spent in MVPA was used for the physical activity level. MVPA is defined as physical activity levels in SOFIT of 4 (walking) or 5 (running; 15). For CARS, the mean interval score was used for the activity score. No conversion to MVPA was available for CARS. The data were analyzed separately for each instrument using a completely crossed 2 by 3 (trial by rater) ANOVA with all participants’ activity coded on two trials by three raters. Seven sources of variability were estimated from the ANOVA results. The variance associated with participant (σ²_p), trial (σ²_t), rater (σ²_r), three two-way interactions, and the residual term (three-way interaction plus error) were determined using the VARCOMP procedure from SAS, version 8 (16). Negative variance components were set to zero in subsequent calculations as suggested by Morrow (11).

The participant variance component (σ²_p) indicates how much the children differed in their level of physical activity during program activities. It is assumed that the majority of the variability in scores should be due to the true differences among the participants being observed. Sampling variability due to trial (σ²_t) gives an indication of sources of error due to the test–retest reliability of the raters. Sampling variability due to the rater (σ²_r) gives an indication of interrater reliability. Three two-way interaction terms were calculated. The person-by-trial (σ²_p × σ²_t) interaction gives an indication of how the trials were influenced by the individual activity patterns (i.e., were between-participant differences consistent across trials?); since both trials of the child were captured on videotape, this interaction term should be close to zero in this study. Variability due to person-by-rater interaction (σ²_p × σ²_r) provides information regarding whether a rater’s coding is influenced by individual activity patterns. The rater-by-trial (σ²_r × σ²_t) interaction provides information on how the raters are influenced by coding the videotapes on two occasions (i.e., were rater differences consistent across trials?). The residual variance component (three-way interaction term) will include systematic and random error (17). Percentage of variances were calculated by dividing the variance for each component by the total variance.

Generalizability coefficients (g-coefficients) were calculated using variance components to give an indication of reliability. G-coefficients give an indication of how sound the generalization for an individual’s observed score is to the person’s true universe score. G-coefficients are calculated differently depending on what decisions are to be made from the results. If interpretations pertain to the relative standing of participants, then a relative decision is made. G-coefficients for relative decisions are calculated using all the interaction terms between the facets and the object of measurement. If decisions are based on the absolute score for participants and relative standing has no bearing, then absolute decisions are made. In the case of absolute decisions, all variance components except the object of measurement are used in the calculation of an index of dependability. The index of dependability is the absolute decision equivalent of the g-coefficient in relative decisions (17). Because relative standing of the individuals in the study had no bearing on the results, estimates for measurement error were based on absolute interpretations.

Validity coefficients were calculated using Pearson product moment correlations between activity counts from Actiwatch accelerometers and the results from the two systemic observation tools. The correlation was conducted with all interval observation scores for all participants’ intervals and all accelerometer counts.
Results

According to SOFIT, children with MR participated in MVPA an average of 26.48% ($SD = 14.77\%$) of the programming time. CARS indicates that the overall mean activity level of the children was 2.37 ($SD = 0.35$) on a scale of 1 to 5, where level 1 is stationary with no limb movement and level 5 is very fast translocation. Table 1 summarizes each child’s average percentage of MVPA across the six trials during the physical activity program from SOFIT and each child’s physical activity level across three raters and two trials as coded by CARS.

**SOFIT**

Variability of MVPA for SOFIT was examined using a participant-by-rater-by-trial G-study. True differences in physical activity levels are expected and, therefore, the participant variable was the largest source of variance in MVPA level (94.23%). The second largest source of variance for SOFIT was the residual term, which was approximately 3% of total variance, indicating that there was little random measurement error. No other facets or interactions had high proportions of variance, which indicates that trial, rater, and two-way interaction terms were associated with a relatively low level of measurement error. The index of dependability ($\Phi$) for SOFIT was 0.98, indicating there is strong evidence of the ability to generalize the children’s observed MVPA scores to their universe score. Table 2 shows variance component estimates for SOFIT. Validity evidence was determined using Pearson product moment correlations. A correlation was calculated using scores from 15 s momentary time and accelerometer counts in 15 s intervals for all participants. A correlation between the results of SOFIT and accelerometer counts was $r = .10$.

<table>
<thead>
<tr>
<th>Participant</th>
<th>SOFIT</th>
<th></th>
<th></th>
<th>CARS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>52.97</td>
<td>4.99</td>
<td>2.84</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>29.93</td>
<td>2.94</td>
<td>2.25</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>34.85</td>
<td>5.61</td>
<td>2.93</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>36.39</td>
<td>3.23</td>
<td>2.52</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>14.38</td>
<td>1.43</td>
<td>2.17</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>43.60</td>
<td>2.96</td>
<td>2.34</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>30.38</td>
<td>1.50</td>
<td>2.33</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5.60</td>
<td>1.16</td>
<td>2.04</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>10.83</td>
<td>2.04</td>
<td>2.29</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>15.18</td>
<td>1.51</td>
<td>2.12</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>16.92</td>
<td>7.53</td>
<td>2.27</td>
<td>0.11</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1** Descriptive Statistics for Children’s Percentage of MVPA for SOFIT and Children’s Activity Scores for CARS

*Note:* Means are based on the average score across the three raters (six trials).
Variance estimates of activity scores (mean interval scores) were computed for CARS. As expected, the largest variability was due to person, with a variance of 49.65%. For CARS, the highest sources of variance after participants were the rater facet (31.49%) and the participant-by-rater interaction term (15.41%). Since terms including rater account for a high proportion of variance, there is an indication that there were discrepancies between the raters’ coding of participants’ physical activity levels. Contributions of the trial and other interaction terms to the total variance were negligible. This indicates that trial and residual errors do not contribute significantly to the total variance. With the testing protocol used (3 raters × 2 trials), CARS had Φ of 0.75. This provides moderate evidence that the children’s observed activity scores may be generalized to their universe score. Correlation between criterion measurement and activity scores from CARS was r = .61. Table 3 summarizes variance component estimates for CARS.

**Discussion**

The results of the present study provide evidence that by using three raters and two trials, SOFIT has high reliability (Φ = 0.98). This provides strong evidence that SOFIT coding scores may be generalized to the individual’s true universe score. Error variances due to trial, rater, and the trial-by-rater interaction were all low (0.30%, 0.67%, and 0.44%), which indicate that different raters and trials are not sources of error for SOFIT. These small variances indicate that the training provided for SOFIT in this study was adequate to code physical activity patterns of children with MR. No previous studies have used G-theory to investigate the psychometric properties of SOFIT, however; the low error associated with raters lends weight to previous studies that suggest SOFIT had high interrater reliability (9). As a follow-up calculation, indices of dependability were determined using a decision study (d-study) to determine the optimal measurement protocol (see

<table>
<thead>
<tr>
<th>Variation</th>
<th>Sum of squares</th>
<th>Df</th>
<th>Mean squares</th>
<th>Estimated variance components</th>
<th>Relative magnitude*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants (p)</td>
<td>13634</td>
<td>10</td>
<td>1363.42</td>
<td>224.66</td>
<td>94.23%</td>
</tr>
<tr>
<td>Rater (r)</td>
<td>125.62</td>
<td>2</td>
<td>62.81</td>
<td>1.60</td>
<td>0.67%</td>
</tr>
<tr>
<td>Trial (t)</td>
<td>39.33</td>
<td>1</td>
<td>39.33</td>
<td>0.71</td>
<td>0.30%</td>
</tr>
<tr>
<td>p × r</td>
<td>189.54</td>
<td>20</td>
<td>9.48</td>
<td>0.41</td>
<td>0.17%</td>
</tr>
<tr>
<td>p × t</td>
<td>119.68</td>
<td>10</td>
<td>11.97</td>
<td>1.99</td>
<td>0.83%</td>
</tr>
<tr>
<td>r × t</td>
<td>19.64</td>
<td>1</td>
<td>19.64</td>
<td>1.06</td>
<td>0.44%</td>
</tr>
<tr>
<td>Residual (p × r × t,e)</td>
<td>167.73</td>
<td>21</td>
<td>7.98</td>
<td>7.99</td>
<td>3.3%</td>
</tr>
</tbody>
</table>

*Relative magnitude was calculated using estimated variance divided by the total variance.
Table 3  Variance Component Estimates and Their Relative Magnitude for CARS

<table>
<thead>
<tr>
<th>Variation</th>
<th>Sum of squares</th>
<th>Df</th>
<th>Mean squares</th>
<th>Estimated variance components</th>
<th>Relative magnitude %**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants (p)</td>
<td>4.781</td>
<td>10</td>
<td>0.478</td>
<td>0.0721</td>
<td>49.65</td>
</tr>
<tr>
<td>Rater (r)</td>
<td>2.114</td>
<td>2</td>
<td>1.057</td>
<td>0.0457</td>
<td>31.49</td>
</tr>
<tr>
<td>Trial (t)</td>
<td>0.002</td>
<td>1</td>
<td>0.002</td>
<td>0*</td>
<td>0</td>
</tr>
<tr>
<td>p x r</td>
<td>0.983</td>
<td>20</td>
<td>0.049</td>
<td>0.0224</td>
<td>15.41</td>
</tr>
<tr>
<td>p x t</td>
<td>0.025</td>
<td>10</td>
<td>0.003</td>
<td>0*</td>
<td>0</td>
</tr>
<tr>
<td>R x t</td>
<td>0.007</td>
<td>1</td>
<td>0.007</td>
<td>0.0002</td>
<td>1.46</td>
</tr>
<tr>
<td>Residual (p x r x t,e)</td>
<td>0.101</td>
<td>21</td>
<td>0.005</td>
<td>0.0048</td>
<td>3.31</td>
</tr>
</tbody>
</table>

*Negative variance components were set to zero; **relative magnitude was calculated using estimated variance divided by the total variance.

Table 4  Decision Study for SOFIT (p x r x t design)

<table>
<thead>
<tr>
<th>Number of raters</th>
<th>Number of trials</th>
<th>Phi coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.98</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.96</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Table 4). SOFIT maintained a high level of reliability \( \Phi = 0.94 \) with only one rater and trial.

The results of the present study provide evidence that CARS has moderately high dependability \( \Phi = 0.75 \). This index of dependability is similar to intracorrelation coefficient value. This provides moderate evidence that CARS scores can be generalized to true universe scores. Error variance due to rater (31.49%) and the participant-by-rater interaction (15.41%) accounted for the largest proportions of variance other than participants, which indicates the rater facet may need to be adjusted to achieve a higher level of generalizability. Raters received 7 hr of training on the two instruments with interrater agreement of 87% to 93%. Puhl et al. (14) provided raters with 30 hr of training with CARS over an 8-week period prior to data collection and interrater reliability estimates.

The low variance associated with trials and the interaction terms including trials indicate that trial does not contribute as a source of error in this study. Indices of dependability were also calculated for CARS using a decision study (d-study) to determine the optimal measurement protocol (see Table 5). With the level of training provided in this study, CARS would need four raters with one trial to reach a dependability level of 0.80, which is a commonly recommended level of reliability coefficient (17). Using only one rater and trial would not be a sound decision with
the low index of dependability ($\Phi = 0.50$). Given the low indices of dependability with one rater, CARS would require a more extensive training protocol before being implemented with a single rater.

The validity evidence for SOFIT was not strong. The low correlation indicated that there is almost no association between SOFIT measurements and accelerometer counts. This result differs from that by Rowe et al. (15), who provided evidence for the validity of SOFIT levels citing significant differences in heart rates between SOFIT levels. Results of the present study and the study conducted by Rowe et al. (15) may differ because the studies were conducted in different settings. Rowe’s study was conducted by having the participants spend 5 continuous minutes at each SOFIT level and collecting heart rate data. As Horvat and Franklin (6) suggested when measuring physical activity of individuals with MR, “physical activity must be measured in the natural environment to obtain an accurate assessment” (p. 189). The present study was conducted in an authentic activity environment.

Another possible explanation for a lower validity coefficient is that individuals with MR are more sedentary than their peers without MR (1), and SOFIT is unable to differentiate among the different levels of physical activity for children with MR. The data indicate that the participants in this study did have low levels of physical activity. SOFIT had an average MVPA of 26.47% ($SD = 14.77\%$). The inability of the scales to differentiate between the lower physical activity levels leads to a reduction in the variability and can deflate correlations. One noticeable difficulty with SOFIT is that it does not differentiate between walking speeds, so when a child is walking at a very slow pace, he or she is coded at level 4, but if another child is walking very quickly, he or she is also coded at level 4. For example, the results of descriptive statistics from SOFIT indicated that children with MR spent most of the time standing (Level 3, 52% of the time) or walking (Level 4, 25% of the time), whereas the results of CARS indicated that 16% of the time children walked at a very slow speed and 61% of the time children were stationary with some movement. This inability to discriminate low- and high-intensity exercise may contribute to the lower validity coefficient for SOFIT.

Another potential explanation for the lower validity coefficient for SOFIT is that accelerometers with hand placement may be an inappropriate criterion measure for physical activity levels. The lessons in which activity was recorded were sometimes sedentary in nature with a great deal of arm movement (e.g., activities involving rhythmic gymnastics ribbons and parachutes). This may have resulted in a high activity count on the accelerometer because the accelerometer was placed
on the child’s arm. It is possible that a child may have been seated while shaking a parachute, which would be a level 2 activity in SOFIT, but the accelerometer would register a great deal of movement because the arm was constantly moving. An overestimation of physical activity levels by the accelerometer would provide an explanation for the lower correlation.

CARS provided somewhat stronger evidence than SOFIT for validity. A correlation was calculated using each mean interval score and interval accelerometer count for all participants ($r = 0.61$). Although this validity coefficient was not strong as a recommended level, it has met the minimal recommended validity coefficient level. Validity evidence was reported by Puhl et al. (14), who cited significant differences in VO$_2$ uptake at the different CARS levels.

The results of this study indicate that SOFIT and CARS can reliably generalize activity levels from an observed score to a true universe score for individuals with MR. However, the results of this study also illuminate concerns about the validity evidence of systematic observation tools in a natural activity setting. While previous studies have demonstrated that SOFIT and CARS provide good evidence of validity, those studies have all been in artificial activity environments. Based on the findings in this study, validity evidence for SOFIT may not generalize to individuals with MR. It appears that SOFIT cannot discriminate between the physical activity levels of the participants in this study. CARS demonstrates stronger validity evidence than SOFIT, but it is important to have sufficient rater training before using CARS for measuring physical activity level of individuals with MR.

References


