Feasibility and Reliability of Tests Measuring Health-Related Physical Fitness in Children With Moderate to Severe Levels of Intellectual Disability

Marieke Wouters, Anna M. van der Zanden, Heleen M. Evenhuis, and Thessa I. M. Hilgenkamp

Abstract
Physical fitness is an important marker for health. In this study we investigated the feasibility and reliability of health-related physical fitness tests in children with moderate to severe levels of intellectual disability. Thirty-nine children (2–18 yrs) performed tests for muscular strength and endurance, the modified 6-minute walk test (6mwt) for cardiorespiratory fitness, and body composition tests, and 30–97% of the tests were successfully completed. Short-term test-retest reliability of all tests was good (Intraclass Correlation Coefficient [ICC] \( > .8 \)), long-term test-retest reliability was good for most tests (ICC \( > .7 \)), but low ICCs were found for most strength tests. Measuring body composition and cardiorespiratory fitness is feasible and reliable. Measuring muscle endurance is fairly feasible and reliable.

Key Words: intellectual disability; children; physical fitness tests; reliability

High levels of physical fitness are associated with lower risks for a range of negative health outcomes in children and adolescents (Hurtig-Wennlof, Ruiz, Harro, & Sjostrom, 2007; Ortega, Ruiz, Castillo, & Sjostrom, 2008; Ortega et al., 2007; Smith et al., 2014). Children with intellectual disability (ID) may have more chronic health conditions than children who develop typically (Oeseburg, Dijkstra, Groothoff, Reijneveld, & Jansen, 2011), and their physical fitness seems to be lower than that of typically developing children as well (Gillespie, 2003; Golubovic, Maksimovic, Golubovic, & Glumbic, 2012; MacDonncha, Watson, McSweeney, & O’Donovan, 1999; Mercer & Lewis, 2001). Measuring physical fitness is important to identify children in need of interventions and to evaluate these interventions. Moreover, objective information on the level of fitness of these vulnerable children can be useful for care organizations to set priorities for policy and practice.

However, measuring physical fitness in this population is challenging. Limitations in intellectual functioning and communication, sensory impairments, limb malfunctions, delayed growth and motor development, autism spectrum disorder and challenging behavior, and lack of understanding or motivation to perform the test with the required complete exertion or maximum performance (Frey & Chow, 2006; Halle, Gabler-Halle, & Chung, 1999; Hartman, Houwen, Scherder, & Visscher, 2010; Shea, 2006; Simons, Daly, Theodoroe, Caron, & Andoniadou, 2008; Westendorp, Hartman, Houwen, Smith, & Visscher, 2011; Westendorp, Houwen, Hartman, & Visscher, 2011) can all influence test performance. Field-based tests are often the tests of choice for measuring physical fitness in this population (Fernhall et al., 1998; MacDonncha et al., 1999; Pizarro, 1990). These tests are cost- and time-efficient and can be conducted in a familiar environment for the participants, in contrast to a test conducted in a laboratory. Lab-based tests provide more objective and accurate measurements, therefore, it is important to establish the psychometric properties of field-based tests. In our
literature review, we studied the properties of tests in children with ID and concluded that feasibility, reliability, and validity are not well-established in children with ID (Wouters, Evenhuis, & Hilgenkamp, 2017). Furthermore, younger children and children with more severe levels of ID were underrepresented in published studies, which limits the generalizability of the results and the evidenced-based application of physical fitness tests in prevention and treatment of chronic disability or disease in all children with ID.

Therefore, potentially suitable tests should be studied to determine their feasibility and psychometric properties in young children and children with more severe levels of ID. Based on the literature review conducted by the authors (Wouters et al., 2017) and focus groups, we selected body mass index (BMI) and waist circumference for body composition, and four tests from the Functional Strength Measurement (FSM) for muscular strength and endurance. The FSM is a test battery developed and tested in children aged 4–10 years (Aertssen, Ferguson, & Smits-Engelsman, 2016). The six-minute walk test (6mwt) was modified to measure cardiorespiratory fitness. For these selected tests, positive results on feasibility, reliability, and/or validity had been found in adolescents with mild to moderate levels of ID (Bandini, Fleming, Scampini, Gleason, & Must, 2012; MacDonncha et al., 1999; Tejero-Gonzalez et al., 2013) or school-aged children with mild levels of ID (Bullens, 2012), but no information is yet available on the feasibility and psychometric properties in children with moderate to severe levels of ID.

For that reason, this study focuses on the feasibility and short- and long-term reliability of these field-based physical fitness tests in children with moderate to severe levels of ID, in order to determine the suitability of these tests for application in clinical practice and in research.

Materials and Methods

Participants
This study was performed in collaboration with a service provider for people with disabilities in the Netherlands. Three children’s daycare facilities, in which 130 children received daycare support, were selected.

We invited all children aged 2–18 years with a moderate or severe level of ID and sufficient motor capacities to walk independently to participate. The first inclusion criteria (level of ID) was evaluated by the behavioral therapist or psychologist of the child by reviewing existing testing results from the Bayley Scale of Infant and Toddler Development, Third Edition (BSID-III; Bayley, 2006), the Snijders-Oomen Nonverbal Intelligence Test (Tellegen, Winkel, Wijnberg-Williams, & Laros, 1998), or comparable tests. Children were invited if their testing results corresponded to an IQ of 20–55. The second inclusion criteria (motor capacity) was evaluated by the physical therapist of the child by answering the question, “Can this child walk independently?” After this first screening, parents or legal representatives of the children who met the inclusion criteria received an invitation letter. Children were not invited if the behavioral therapist, psychologist, or the medical staff advised against participation in the study because of a negative influence on the behavior or physical health of the child (exclusion criteria).

Seventy-one of 130 children (55%) were invited to participate. Reasons for not inviting were a mild or profound level of ID (n = 20), insufficient motor capacities (n = 9), or both (n = 16). One child could not be invited because of contra-indications due to a severe autism spectrum disorder (ASD), and 13 children were not present during the testing period. For 41 children (58%), we received informed consent and they were enrolled in the study (Figure 1). The parents or legal representatives of these children gave informed consent and completed the Physical Activity Readiness Questionnaire (parQ) for further medical clearance. The included participants were not different than the invited nonparticipants on demographics (age and sex, data not shown).

Ethical approval was obtained (MEC-2013-491) from the ethics committee of the Erasmus Medical Center. The study adheres to the Declaration of Helsinki for research involving human subjects (World Medical Association, 2013).

Measurements
Information on the level of ID, autism spectrum disorder, and behavioral problems was provided by the behavioral therapist or psychologist of the participants. Information on age, Down syndrome, and physical problems was extracted from the medical file. The gross motor scale of the BSID-III (Bayley, 2006) was completed by physical therapists and was used to give insight into their
motor development. Because this test was developed for children up to 42 months, raw scores were not converted into standardized norm scores. We used the raw scores to check if completion rates were significantly different for relative low and high motor development.

**Physical fitness tests.** Health-related physical fitness can be divided in five measurable components: body composition, muscular strength, muscular endurance, cardiorespiratory fitness, and flexibility (American College of Sports Medicine, 2014). Flexibility has not been included in this study, because the relationship between flexibility and health has not been confirmed in children (Ganley et al., 2011; Ruiz et al., 2009). Potentially suitable tests were selected based on our literature review (Wouters et al., 2017), with the addition of extra muscular strength and endurance tests tested in other pediatric populations. All tests were discussed with a focus group of 10 physical therapists working with individuals with ID for more than 3 years. In the final selection, some tests with positive psychometric properties found in the review were eliminated, based on the experience and estimation of the focus group.

**Body composition.** Body mass index (BMI) was calculated as body weight in kilogram divided by height in meters squared. The participant was on bare feet and wore light clothes. Height was measured with a portable stadiometer accurate at 0.1 cm level. The participant was asked to stand against the stadiometer. The participant had to stand against, with feet flat on the ground and face in the “Frankfurt plane,” with the lower margin of the left orbit and the upper margin of each ear canal at the same plane. Body weight was measured using an electronic calibrated scale accurate at 0.1 kg level. The participant had to stand independently and motionless for at least 2 seconds to properly measure the weight.

The BMI scores were compared to the World Health Organization (WHO) growth references (de Onis et al., 2007; WHO Multicentre Growth Reference Study Group, 2006). The WHO classification for underweight, overweight, and obesity is based on BMI-for-age z-scores. BMI $-2SD$ is classified as underweight. Children at age 0–5 years with BMI $+2SD$ are described as overweight, and $+3SD$ as obese. For older children, $+1SD$ is classified as overweight and $+2SD$ as obese (de Onis & Lobstein, 2010).

Waist circumference provides a measure of abdominal adiposity (Taylor, Jones, Williams, & Goulding, 2000) and was measured using a flexible measuring tape accurate at the 0.1 cm level. The participant was asked to stand still, with the arms down. The circumference was measured halfway between the iliac crest and tenth rib, directly on the skin. This measurement was found reliable (ICC = .98) in 17 adolescents.
Muscular strength and endurance. The muscular strength and endurance tests were selected from the Functional Strength Measurements (FSM; Aertssen et al., 2016). This test battery consists of eight items assessing the functional strength of the upper and lower extremities, and was developed for children 4 to 10 years old. It was studied in children with no disability (Aertssen et al., 2016), and with mild levels of ID (Bullens, 2012). Good test retest reliability (ICC = .7–.9; Aertssen et al., 2016; Bullens, 2012), internal consistency (α = .7; Aertssen et al., 2016), and fair validity (r = .5–.7; Aertssen et al., 2016) were found. We selected three throwing tasks and the standing long jump for muscular strength and stair climbing for muscular endurance, because these tests were the most functional and, thereby, understandable tests according to the focus group. For all of the muscular tasks, the participant practiced at least once prior to the test starting, with a maximum of five practice attempts prior to the test. Improvement was usually seen between the first and third practice of the task, but, when necessary, participants were allowed to practice up to five times to ensure correct execution or to verify that the participant was not able to execute the task correctly. After the practice session, the participant performed the test three times, with at least 30 seconds rest in between. The maximum score of these three attempts was the final score. For the throwing tasks, different bags were used (1, 2, and 3 kg), and bag selection was based on the age of the participant. Due to the fact that the FSM is developed for children ≤10 years, we added a bag of 4 kg for children of ages 11 and above.

During the overarm throwing task, the participant stood behind a line marked on the ground with the feet slightly apart, raised the bag behind his or her head, and threw the bag as far as possible. The distance from the line to the proximal side of the bag was measured. In the underarm throwing task, the participant stood behind a marked line with legs spread apart. The participant bent forward, while holding the bag, and swung the bag back, between the legs, and then threw it forward. The distance between the marked line and the proximal side of the bag was measured in centimeters.

For the chest pass test, the participant sat on the ground with his or her back against the wall and legs spread on the ground. The participant held the bag in both hands, with the back of his or her hands against the chest and the elbows pointed sideways. The participant passed the bag as far as possible, without bending forwards. The distance between the proximal side of the bag and the wall was measured in centimeters.

For the standing long jump, the participant stood behind a line marked on the ground, with his or her feet slightly apart. The participant jumped as far as possible with both feet together by bending his or her knees and swinging the arms. The distance between the line and the proximal side of the heels was measured in centimeters.

Cardiorespiratory fitness. The six-minute walk test (6mwt) is a test of functional capacity, and is recommended by the American College of Sports Medicine (ACSM) to use as a field-based test for cardiorespiratory fitness in several populations with low fitness levels (American College of Sports Medicine, 2014). During the 6mwt, the participants walk as many meters as possible in 6 minutes. A review of Bartels, de Groot, and Terwee (2013) indicated good test-retest reliability for children with chronic health conditions, and in some of these populations criterion validity was sufficient as well. In adolescents who were overweight and had mild to moderate levels of ID, Elmahgoub, Van de Velde, Peersman, Cambier, and Calders (2012) found good reliability (ICC = .82) and fair validity (r = .69). Better test-retest reliability was found in adults with ID (ICC = .98; Nasuti, Stuart-Hill, & Temple, 2013). They also found good concurrent validity with the maximum oxygen uptake (r = .84). Sufficient concurrent validity was also found in adults with Down syndrome (DS; r = .78; Boer & Moss, 2016).

In the original 6mwt, the participant independently walks 30 meters back and forth in a corridor to reach as many meters as possible in 6 minutes, with standardized encouragement (American Thorax Society, 2002). We modified the test to make it more feasible for children with ID, because more than half of the participants were unable to perform the test without physical guidance (i.e., for direction; unpublished data). Therefore, the in-

M. Wouters et al.
structor gave the participants a loose hand, without pulling along. Furthermore, the participants walked a square of 20 meters and were encouraged individually to keep walking as fast as they could. The total distance covered in 6 minutes was recorded in meters. Before and after the test, the participant sat on a bench. The heart rate was continuously monitored, before (1 minute), during, and after the test (3 minutes). Peak and average heart rate were calculated afterwards.

**Procedure**

The health-related physical fitness of the participants was tested on two occasions, with 2 to 4 weeks between sessions 1 and 2, and the third session within an hour after session 2 (see Figure 2), to study both the long-term and short-term reliability. The three test sessions were conducted at the same location at the same time of the day and tested by the same instructor. Body composition was only measured in session 1 and 2. The instructors were physical therapists or human movement scientists, all experienced in working with children with ID. In a 1–2 hour workshop, all instructors were taught how to conduct the tests according to protocol. An extra instructor was always present to note the scores and check for protocol deviations.

The tests were performed in a gymnasium at the three different daycare facilities during daytime. The tests were performed in a fixed order within each test session: body composition, modified 6-minute walk test, overarm throwing, stair climbing, underarm throwing, standing long jump, and chest pass. In this sequence, we took into account the load on the upper and lower extremities, respectively, and familiarization with the tester and the test environment by starting with relatively easy tests (i.e., body composition and the modified 6mwt). Resting between tests was allowed. The instructors explained the tests and demonstrated how to do them. For the muscular strength tests, the participants could practice up to five times, as previously discussed. When they performed the task correctly during this practice phase, the actual test started. During all tests, the instructors encouraged the participants to perform maximally. No standardized encouragement was used, because every child has their own preferences and communication style. If a test was not performed according to protocol, or not performed at all, the presumed reason for dropout was noted (e.g., cognitive issues, motor development issues, challenging behavior, concentration issues, or issues with logistics). The instructors also indicated the level of effort during the tests. If a participant was unable to perform the test with maximum effort, the test was scored as invalid and the presumed reason was noted. The dropout statistics can be found in Table 2.

**Figure 2.** Test procedure.
Descriptive statistics were used to describe the participants and the feasibility outcomes. Feasibility of the test was based on the completion rate. When completion rates were ≥75% the test was feasible; a test was considered fairly feasible when completion rates were 50–75%, and not feasible when completion rates were <50%. Chi-square tests were used to check if completion rates were significantly different for sex, age groups (2–4; 5–8; 9–12; 13–17 y), level of ID, overweight (zBMI ≤ 1 vs. > 1), diagnosed with ASD, challenging behavior, and motor development groups (raw BSID-III score <62 vs. ≥62). A raw BSID-III score of 62 is comparable to a motor development of a typically developing child of 33 months.

Normality of the fitness parameters was checked by use of a Kolmogorov-Smirnov test. If the parameters were normally distributed, systematic difference in mean scores over time was checked by a paired T-test. The Wilcoxon Signed Rank test was used as the nonparametric alternative.

The test-retest reliability was studied by use of the ICC (two-way mixed model, absolute agreement, single measures). The ICC provides insight into the relative reliability by comparing the within-subject variability to the between-subject variability (Atkinson & Nevill, 1998). An ICC > .7 was indicated as sufficient test-retest reliability (De Vet, Terwee, Mokkink, & Knol, 2011).

Measurement error was calculated to give information on the absolute reliability: the degree to which repeated measurements vary for individuals (Atkinson & Nevill, 1998). The Standard Error of Measurement (SEM) estimates the standard error in a set of repeated scores. It is calculated by $SEM = SD \times \sqrt{1-ICC}$, where $SD$ is the pooled standard deviation of the measurements. In a reliable measurement, the SEM would be small. The Minimal Detectable Change (MDC) at 95% CI is defined as the minimum change that falls outside the measurement error in the score of an instrument and facilitates the interpretation of changes in test scores. It is calculated as $MDC = 1.96 \times SEM \times \sqrt{2}$. The SEM and MDC are depicted in the measurement unit of the test (De Vet et al., 2011).

An A Bland-Altman plot is a graphical method to compare two test results (Bland & Altman, 1986) and gives insight into the relationship between averages and differences, systematic biases, and possible outliers. The difference between the two tests is plotted against the average of the two tests. Horizontal lines are plotted at the mean difference and at the limits of agreement (LOA), which are defined as mean difference plus or minus 1.96 × standard deviation of the difference. The plots are depicted in the Appendix.

The analyses were performed for session 1 compared to session 2 to establish the long-term reliability, and for session 2 compared to session 3 to establish the short-term reliability. The data was analyzed by use of SPSS 22 (IBM statistics) and alpha was set at 5%.

### Table 1
**Characteristics of the Participants (N=37)**

<table>
<thead>
<tr>
<th>Sex</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>25</td>
<td>68</td>
</tr>
<tr>
<td>Female</td>
<td>12</td>
<td>32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2–4 years</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>5–8 years</td>
<td>13</td>
<td>35</td>
</tr>
<tr>
<td>9–12 years</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td>13–17 years</td>
<td>11</td>
<td>30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level of intellectual disability</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moderate</td>
<td>14</td>
<td>38</td>
</tr>
<tr>
<td>Severe</td>
<td>23</td>
<td>62</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weight class*</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underweight</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Normal weight</td>
<td>23</td>
<td>62</td>
</tr>
<tr>
<td>Overweight</td>
<td>12</td>
<td>33</td>
</tr>
<tr>
<td>Unknown</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Diagnoses</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autism spectrum disorder</td>
<td>14</td>
<td>38</td>
</tr>
<tr>
<td>Down syndrome</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>Challenging behavior</td>
<td>27</td>
<td>73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Motor development</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low (raw BSID-III score &lt; 62)</td>
<td>21</td>
<td>57</td>
</tr>
<tr>
<td>High (raw BSID-III score ≥ 62)</td>
<td>16</td>
<td>43</td>
</tr>
</tbody>
</table>

*Based on BMI-for-age z-scores compared with WHO growth references (de Onis et al., 2007; WHO Multicentre Growth Reference Study Group, 2006).

**Note.** BSID-III = Bayley Scale of Infant and Toddler Development, Third Edition.

Analysis

Descriptive statistics were used to describe the participants and the feasibility outcomes. Feasibility of the test was based on the completion rate. When completion rates were ≥75% the test was feasible; a test was considered fairly feasible when completion rates were 50–75%, and not feasible when completion rates were <50%. Chi-square tests were used to check if completion rates were significantly different for sex, age groups (2–4; 5–8; 9–12; 13–17 y), level of ID, overweight (zBMI ≤ 1 vs. > 1), diagnosed with ASD, challenging behavior, and motor development groups (raw BSID-III score <62 vs. ≥62).
Participants
Initially, 41 children were included, but two participants were not able to participate in the second test day. One participant moved to a different region of the country and changed daycare facilities, and one participant was recovering from a planned operation during the retest. Thirty-nine participants completed the test and retest, but two participants were excluded from analysis: as part of usual care, their level of ID was retested after the data collection and they were evaluated to have a mild level of ID (Figure 1). The characteristics of the remaining 37 participants can be found in Table 1.

Feasibility
The average time to complete the test battery on the first day was 42 minutes (SD = 13), with a range of 12 to 67 minutes. Table 2 depicts the number of successful completions and reasons for dropout in the first session. The body composition tests and the walk test appeared to be feasible with completion rates of 76–97%, and the stair climbing appeared fairly feasible (59%). The muscular strength tests had the lowest completion rates (30–41%) and were considered not feasible for this population.

The most common reason for dropping out was the cognitive abilities of the participants—they did not understand the instructions or aim of the test and were therefore not able to perform the test according to the protocol. In the standing long jump and the stair climbing test, motor development was the limiting factor. The dropout rate was not significantly different for sex; age; level of ID; being overweight; or having DS, ASD, or challenging behavior. However, more participants with low motor development could not perform the muscular strength tests: overarm throwing ($\chi^2 = 8.428; df = 1; p = .004$), underarm throwing ($\chi^2 = 4.985; df = 1; p = .026$), chest pass ($\chi^2 = 5.072; df = 1; p = .029$), and standing long jump ($\chi^2 = 8.397; df = 1; p = .004$).

Test-Retest Reliability
The number of days between the first and second test sessions was, on average, 18 days (SD = 4). Table 3 shows the results of the three sessions and the long- and short-term reliability coefficients (ICC, SEM, MDC). All short-term ICCs (between session 2 and 3) were sufficient (ICC > .7) and the Bland-Altman plots revealed no systematic bias, but some outliers were seen (see the Appendix).

The long-term test-retest reliability (between session 1 and 2) for the body composition measurements were very high (ICC = .99). Of note, the average body weight increased by 0.6 kg over the two sessions ($W = 348.5; p = .048$), but this did not significantly increase BMI. The overarm throwing was sufficiently reliable for the long-term period (ICC = .76), with SEM = 21 cm and MDC = 57 cm. The other muscular strength tests had low ICCs (.55–.64). Moreover, the average standing long jump score increased significantly from the first to the second session, by 11 cm ($p = .035$). The walking stair test, for measuring muscular endurance, had good test-

Table 2
Completion Rates and Reasons for Dropout of the First Session (N=37)

<table>
<thead>
<tr>
<th></th>
<th>Successful completion</th>
<th>Cognitive development</th>
<th>Challenging behavior</th>
<th>Motor development</th>
<th>Concentration</th>
<th>Logistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>35 95%</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Weight</td>
<td>36 97%</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Waist circumference</td>
<td>35 95%</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Overarm throwing</td>
<td>11 30%</td>
<td>16</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Underarm throwing</td>
<td>11 30%</td>
<td>18</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Chest pass</td>
<td>11 30%</td>
<td>18</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Standing long jump</td>
<td>15 41%</td>
<td>8</td>
<td>3</td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Stair climbing</td>
<td>22 59%</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Modified 6mwt</td>
<td>30 81%</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Heart rate monitor</td>
<td>28 76%</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Note. 6mwt = six-minute walk test.
Table 3
Mean Values and Reliability Results

<table>
<thead>
<tr>
<th></th>
<th>Session 1</th>
<th>Session 2</th>
<th>Session 3</th>
<th>Long-term reliability (1 vs 2)</th>
<th>Short-term reliability (2 vs 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>M (SD)</td>
<td>n</td>
<td>M (SD)</td>
<td>n</td>
</tr>
<tr>
<td><strong>Body composition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body mass index (kg/m²)</td>
<td>35</td>
<td>19.5 (5.2)</td>
<td>35</td>
<td>19.6 (5.2)</td>
<td>—</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>35</td>
<td>134 (25)</td>
<td>35</td>
<td>134 (25)</td>
<td>—</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>36</td>
<td>37.9 (23.2)</td>
<td>35</td>
<td>38.5 (23.5)*</td>
<td>—</td>
</tr>
<tr>
<td>Waist circumference (cm)</td>
<td>35</td>
<td>65.4 (18.2)</td>
<td>35</td>
<td>66 (18)</td>
<td>—</td>
</tr>
<tr>
<td><strong>Muscular strength</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overarm throwing (cm)</td>
<td>11</td>
<td>151 (46)</td>
<td>13</td>
<td>151 (46)</td>
<td>13</td>
</tr>
<tr>
<td>Underarm throwing (cm)</td>
<td>11</td>
<td>171 (88)</td>
<td>13</td>
<td>187 (84)</td>
<td>14</td>
</tr>
<tr>
<td>Standing long jump (cm)</td>
<td>11</td>
<td>161 (41)</td>
<td>9</td>
<td>152 (39)</td>
<td>8</td>
</tr>
<tr>
<td><strong>Muscular endurance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stair climbing (n)</td>
<td>22</td>
<td>24 (10)</td>
<td>19</td>
<td>26 (12)*</td>
<td>15</td>
</tr>
<tr>
<td><strong>Cardiorespiratory fitness</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modified 6mwt (m)</td>
<td>30</td>
<td>331 (63)</td>
<td>31</td>
<td>330 (79)</td>
<td>26</td>
</tr>
<tr>
<td>Peak heart rate (bpm)</td>
<td>28</td>
<td>138 (17)</td>
<td>30</td>
<td>133 (16)</td>
<td>24</td>
</tr>
<tr>
<td>Average heart rate (bpm)</td>
<td>28</td>
<td>124 (13)</td>
<td>30</td>
<td>122 (13)</td>
<td>24</td>
</tr>
</tbody>
</table>

Note. 6mwt = six-minute walk test; ICC = Intraclass Correlation Coefficient; SEM = Standard Error of Measurement; MDC = Minimal Detectable Change. *Significant difference between session 1 and 2: weight: W = 348.5; p = .048; standing long jump: t₁₃ = −2.347; p = .035; stair climbing: t₂₀ = −2.222; p = .038.
retest reliability (ICC = .92), but relatively high measurement error (SEM = 3; MDC = 8) and showed significant increase over time between session 1 and 2 ($t_{20} = -2.22; p = .038$). No difference between session 2 and 3 was found. The modified 6mwt was found reliable (ICC = .78), with a SEM of 29m and a MDC of 80m. Measuring the average heart rate during this test was reliable (ICC = .74).

**Discussion**

In this study, we assessed the feasibility and test-retest reliability of nine field-based physical fitness tests in 37 children with moderate or severe levels of ID, with the aim to find appropriate tests to use in this specific population. Body composition (BMI, waist circumference) and cardiorespiratory fitness (6mwt with modifications) can be measured feasibly and reliably, whereas muscular strength and strength endurance tests are hardly feasible and hardly reliable in this group. Only overarm throwing (strength) and stair climbing (endurance) had sufficient long-term reliability.

The reliability coefficients found for body composition measurements are comparable with those in typically developing children (Berkson et al., 2013; Lohman, Roche, & Martorell, 1988), adolescents with mild to moderate levels of ID (MacDonncha et al., 1999; Tejero-Gonzalez et al., 2013), and adults with severe levels of ID and sensory disabilities (Waninge, van der Weide, Evenhuis, van Wijck, & van der Schans, 2009). We found a small, but significant, increase of weight between sessions (0.6kg), not effecting the reliability of BMI. The weight increase is likely to be caused by the day-to-day variation within an individual (Khosla & Billewicz, 1964), or difference in the clothes they were wearing. The maximum difference was 1.5 kg, and increases as well as decreases in body weight were seen (see Bland-Altman plot 3.A, Appendix).

For muscular strength, low feasibility and reliability were found. The throwing tests were hard to understand for the participants (cognitive issues were the main reason for dropping out), and participants with a higher motor development were more likely to perform the test according to protocol. As these tests ask a lot of motor coordination and task understanding, it is not surprising that we found lower reliability than the values found in typically developing children or children with mild levels of ID. In these populations, ICCs of .7-.9 were found (Bullens, 2012; Aertssen et al., 2016). Only the overarm throwing task had sufficient reliability, but the SEM was large (23 cm) and the completion rate low (30%). It is possible that throwing a bag is not an activity that these children are used to. We believe that practicing these tasks will eventually increase the feasibility. The motor skills will develop, and task understanding could possibly be improved in part of the group of children. However, if we want feasible tests to be applicable in a wide population, tests need to be easy to understand and perform, and not require much practice and familiarization. Because there is no alternative test that has been shown feasible and reliable in children with moderate to severe levels of ID, overarm throwing can be used for children that can perform this task until a more suitable test is found.

The reliability of the walking stair test was sufficient and comparable with studies in children with no ID (Aertssen et al., 2016) and children with mild levels of ID (Bullens, 2012). Stair climbing is a functional task that children often perform during their activities of daily living and, therefore, the task understanding was supposed to be greater. In this sample, the score on the walking stair test increased significantly between session 1 and 2, and not between session 2 and 3. Therefore, we recommend practicing the whole test with children at least one time before scoring. A limitation of this test is that children with low motor development cannot perform this test, as can be seen in the results.

The 6mwt was modified to make it more feasible for children with moderate to severe levels of ID. The results we found are comparable to the results found in adolescents who were overweight and had mild to moderate levels of ID (ICC = .82, SEM = 30m, MDC = 83m; Elmahgoub et al., 2012). In typically developing children (Li et al., 2005), in children with a chronic health condition (Bartels et al., 2013), and in adults with ID (Nasuti et al., 2013), ICCs of .84-.98 were found.

We measured heart rate during the 6mwt with two purposes. For safety purposes, we monitored heart rate during the test to check for unusual high or low heart rates. Secondly, the achieved heart rate gives a rough estimation of the amount of effort during the test. The 6mwt is developed as a submaximal test (American Thorax Society, 2002), but people with low physical fitness or disease can...
reach near-maximum levels (Jehn et al., 2009). The peak heart rates we found were submaximal and comparable to those found in typically developing children, where average peak heart rates of 136–153 bpm were found during the 6mwt (Lammers, Hislop, Flynn, & Haworth, 2008; Li et al., 2007; Nixon, Joswiak, & Fricker, 1996), indicating that the amount of effort is comparable to these studies. One study demonstrated lower peak heart rates in typically developing children, with an average of 112 bpm during the 6mwt (Priesnitz et al., 2009).

The modifications we made in the 6mwt were successful in decreasing the dropout rate for this specific pediatric group, but they are likely to influence the walking distance. Walking on a continuous track (Bansal et al., 2008; Sciurba et al., 2003) and different task instructions (Weir et al., 2013) are known to increase the test score. Holding the hand of the pacer can increase stability and confidence, and it is possible that participants adapted their walking speed to that of the tester, even though the testers were instructed not to pull them along. Therefore, the results of the modified 6mwt are more likely to overestimate than underestimate the cardiorespiratory fitness of the child. To date, no studies have studied the effect of walking with a pacer, although this has been done more often in individuals with ID (Lavay, Reid, & Cressler-Chaviz, 1990; Nasuti et al., 2013; Oppewal, Hilgenkamp, van Wijck, & Evenhuis, 2013). Ideally no modifications are made to the original test protocol, in order to enable comparison between studies. Yet, when modifications are inevitable, clear description of the adaptations and consistent use of the modification over time is important.

**Study Limitations**

The study sample was a heterogeneous group with an age range of 2 to 17 years. The heterogeneity was also seen in the wide range of scores on the tests, which inflates the ICC results, because the ICC compares the variance between subjects with the variance within subjects. On the other hand, because the target population is a heterogeneous group, this study sample gives a good reflection of the actual situation and improves generalizability of the study results. The ICCs might, therefore, be less suitable for this study than the SEM and MCD.

Due to the high dropout rates in the muscular strength tests, the reliability results are based on small sample sizes. Because of this, the power of the results to detect differences was decreased and estimation of the ICCs less precise (e.g., large confidence interval).

All tests were performed in a fixed sequence. In our sequence, we took into account the load on upper and lower extremities and familiarization with the tester and the test environment by starting with the easy tests of body composition and the modified 6mwt. The sequence of tests can be of influence on the results (i.e., at the end of the test session, energy levels, motivation, or concentration can be reduced). In our results, the reliability of the muscular strength tests at the end of the session was lower, but the dropout rate was the same.

**Directions for Future Research**

We indicated an ICC score of .7 to be sufficient for group analysis. However, for individual analysis, an ICC of .9 is required (Nunnally & Bernstein, 1994). Therefore, and because of the high MDCs we found, we recommend using the selected tests only for group analysis, and not for individual evaluation at this point. Further research is required in subgroups to find out if the tests can be used in specific groups based on age or level of ID.

The validity of the selected tests has not been assessed in children with moderate to severe levels of ID yet. The assessment of criterion validity will be difficult, because gold standard tests are often complicated lab-based tests. Possibly, construct validity is a feasible alternative.

**Conclusion**

Measuring physical fitness of children with moderate to severe levels of ID with field-based tests is feasible, and can be reliably measured for the purpose of group analysis by using BMI, waist circumference, overarm throwing, stair climbing, and the modified 6-minute walk test.

**References**


Received 2/26/2016, accepted 1/3/2017.

We would like to thank all the participants and their parents, the staff of the daycare centres, and the involved physical therapists of Reinaerde.

Authors:
Marieke Wouters, Reinaerde and Erasmus Medical Center (Netherlands); Anna M. van der Zanden, Reinaerde; Heleen M. Evenhuis and Thessa I. M. Hilgenkamp, Erasmus Medical Center.

Correspondence concerning this article should be addressed to Marieke Wouters, Reinaerde, Europalaan 310, 3526 KS, Utrecht, the Netherlands or Marieke Wouters, Erasmus Medical Center, Department of General Practice, Intellectual Disability Medicine, Wytemaweg 80, 3015 CN Rotterdam, the Netherlands (e-mail: m.wouters.2@erasmusmc.nl).
Appendix

Bland-Altman Plots of Long-Term (A) and Short-Term (B) Results

1.A  Body Mass Index (kg/m2)

2.A  Height (cm)

3.A  Weight (kg)

4.A  Waist circumference (cm)
Reliability Fitness Tests in Children

8.A Standing long jump (cm)

8.B Standing long jump (cm)

9.A Stair climbing (n)

9.B Stair climbing (n)

10.A Modified 6-minute walk test (m)

10.B Modified 6-minute walk test (m)