Reliability and Validity of the Test of In-Hand Manipulation in Children Ages 5 to 6 Years

Karina Pont, Margaret Wallen, Anita Bundy, Jane Case-Smith

KEY WORDS
• hand
• outcomes assessment
• pediatrics
• psychomotor performance
• task performance and analysis

The Test of In-Hand Manipulation (TIHM; Case-Smith, 2000) is a five-task test that uses a 9-hole pegboard to examine 2 key components of in-hand manipulation: rotation and translation with stabilization. The authors used Rasch modeling to examine the TIHM’s construct validity, interrater reliability, and test–retest reliability in 45 typically developing children ages 5.5 years to 6.5 years. A version of the test, revised using Rasch modeling, was found to have evidence for adequate construct validity and excellent interrater reliability. However, test–retest reliability over a 2-week retest period was not supported. The THM demonstrates potential as a clinically useful assessment of in-hand manipulation. The test does not examine all aspects of in-hand manipulation, however, and it may have limited sensitivity to the performance of finger-to-palm and palm-to-finger translation. Further validation of the test is needed before the TIHM can confidently be used in occupational therapy practice.


Successful participation in childhood occupations relies on dextrous hand use. The accomplishment of many daily activities requires the manipulation of an object within the hand after grasp to prepare the object for voluntary release or use in the hand. Exner (1989) termed this fine motor skill in-hand manipulation (IHM) and outlined three main types: (1) translation, which describes the movement of an object between the fingertips and the palm of the hand, for example, picking up coins and placing them into a purse; (2) shift, which is characterized by the linear movement of an object at the fingertips, for example, threading a needle; and (3) rotation, in which an object is rotated about one or more of its axes at the fingertips, for example, unscrewing a jar or turning a pen end over end to position it for writing. Another component of IHM is stabilization, in which one or more objects or parts of objects are held in the hand or the palm so that the thumb and radial fingers can participate in another hand skill, for example, storing one or more coins in the palm of the hand while picking up another (Exner, 1992).

IHM skills develop in children between ages 18 months and 7 years (Exner, 1992). Between ages 3 and 6 years, IHM skills develop rapidly; the consistency and maturity of methods used to manipulate objects increases, and the time required to complete tasks decreases (Pehoski, Henderson, & Tickle-Degnen, 1997a, 1997b). No significant difference appears to exist between boys and girls in the performance of IHM skills (Pehoski et al., 1997a, 1997b).

IHM is considered to be an essential component of fine motor skills (Exner, 1989; Smith-Zuzovsky & Exner, 2004; Ziviani & Wallen, 2006). This view is reflected in the literature by several methods developed to measure IHM, none of which have been made readily available for use in clinical practice. The four existing methods described in the literature include the In-Hand Manipulation Test–Quality Section (Breslin & Exner, 1999), the Test of In-Hand Manipulation (TIHM; Case-Smith, 1996), Pehoski et al.’s (1997a, 1997b) unnamed test, and the Observation Method.

A positive relation between IHM and the performance of functional activities, such as handwriting, use of scissors and cutlery, and buttoning, has been widely hypothesized (Breslin & Exner, 1999; Exner, 1990, 1992; Jewell & Humphry, 1993). Little empirical evidence, however, exists to support the proposed hypotheses. One reason for this lack of evidence is conjectured to be the absence of a readily available test of IHM with established psychometric properties. In this study, therefore, we evaluate the psychometric properties of the only available assessment of IHM, the TIHM (Case-Smith, 1996).

Test of In-Hand Manipulation

The TIHM (Case-Smith, 1996) assesses translation with stabilization and rotation and was designed for 3- to 6-year-old children. Using a nine-hole peg board, the five-task test was adapted from the protocol used by Pehoski et al. (1997a, 1997b). The TIHM is simple to complete and takes 5 to 7 min to administer and score. During the first task, the child is instructed to use the fingertips to turn five “men” 180° onto their heads, replacing each in its original hole. Tasks 2 to 5 are translation with stabilization tasks that involve the child picking up two, three, four, or five pegs, respectively, in his or her fingertips; “hiding” the pegs in the palm of the hand; and then replacing them in the pegboard. The time taken to complete the test, the number of times a peg was dropped or stabilized on an external surface, and the quality of IHM skills are recorded.

Although no study has specifically examined the psychometric properties of the TIHM, several characteristics of the assessment potentially contribute to its reliability. These characteristics include standard administration and scoring procedures and objective scoring criteria (e.g., timed tasks, counting drops of a peg). Several studies have demonstrated a relation between aspects of IHM and fine motor performance, which may provide some evidence to suggest the validity of the TIHM. For instance, Case-Smith and colleagues found correlations of various aspects of the TIHM using only time and drop-stabilization scores with the following tests: Minnesota Handwriting Test (Reisman, 1995), high correlations ($r = -0.770$ and $-0.798$, $p < .01$; Cornhill & Case-Smith, 1996); Fine Motor component of the Peabody Developmental Motor Scales (Folio & Fewell, 1983), moderate to high correlations ($r = -0.649$ to $-0.803$, $p < .005$; Case-Smith, 1995, 2000); and Motor Accuracy test of the Southern California Sensory Integration Tests (Ayres, 1980), low to moderate correlations ($r = -0.428$ to $-0.635$, $p < .05$; Case-Smith, 1995; Cornhill & Case-Smith, 1996). Using a protocol similar to the TIHM, Case-Smith (1991, 1993) also found that the time taken to rotate and translate pegs could be used to distinguish between children with and without fine motor delay ($t[28] = 4.15$ and 4.26, respectively, $p < .001$) and children with and without tactile defensiveness and poor tactile discrimination ($\chi^2 = 8.82–9.62, N = 49, p < .05$).

The TIHM is inexpensive and readily available; it seems to have the most potential to be developed for use in research and clinical practice. However, no consistent means of using the various results obtained from the TIHM exists (e.g., time, drops, stabilizations, quality). Case-Smith (1995, 2000; Case-Smith et al., 1998; Cornhill & Case-Smith, 1996) has used TIHM scores several different ways to make inferences about children’s IHM skills. Moreover, for the TIHM to be used confidently, it is important that it demonstrate adequate reliability and validity.

Study Purpose

We examined the construct validity, test–retest reliability, and interrater reliability of the TIHM using Rasch modeling (Bond & Fox, 2003; Linacre, 2006) and if necessary proposed adaptations so that the test would be valid, reliable, and clinically useful. The research questions were as follows:

- What is the evidence for construct validity of the TIHM?
  - (1) Do data from at least 95% of the items conform to the expectations of the Rasch model?
  - (2) Does a principal-components analysis (PCA) of the residuals suggest that all items work together to define a single underlying construct?

- What is the evidence for the TIHM’s test–retest reliability?
  - More specifically, do data points ±1 standard error from at least 95% of children over a 2-week period overlap?

- What is the evidence for the TIHM’s interrater reliability?
  - That is, do data points ±1 standard error from at least 95% of children, when scored by two or more raters, overlap?

Method

Participants

We recruited children from 1 independent and 4 government-run primary schools from the northern and western metropolitan regions of Sydney, New South Wales, Australia. Children in kindergarten and Year 1 from consenting schools were invited to participate in the study. Eligible children were between the ages of 5.5 years and 6.5 years, typically developing, able to understand and follow instructions given during
testing, and potentially available for two testing occasions 2 weeks apart. We selected this age range as children 5 and 6 years old have reasonably well-developed IHM skills, with most aspects of IHM present to be assessed. The small age range increased the sample’s homogeneity. Exclusion criteria were significant impairment of vision, hearing, motor, or cognitive skills or insufficient understanding of English to complete the test. Ethics approval was obtained from the University of Sydney and the New South Wales Department of Education and Training. The study was carried out according to the ethical guidelines of these organizations. Peat, Mellis, Williams, and Xuan (2001) suggested that a minimum of 30 participants is required to examine test–retest reliability. Moreover, Winsteps (Linacre, 2006), a Rasch analysis program, can reliably analyze small sample sizes with as few as 30 observations. However, we recruited 45 children to allow for attrition between the first and second testing sessions.

Instrumentation

We used the TIHM according to the administration and scoring procedures recommended by Case-Smith (1996). Each task was performed twice. Four aspects of performance were scored during each task: (1) time taken to complete the task (time), (2) number of times a peg was dropped (drops), (3) number of times a peg was stabilized on an external surface (stabilizations), and (4) quality of IHM skills as scored on a 3-point scale (quality; 0 = no IHM skills used, 1 = IHM used less than 50% of the time, 2 = IHM used more than 50% of the time).

Procedure

We conducted a pilot study with 29 children from one school to consolidate administration and scoring procedures. After this study, five different schools participated in testing between May and August 2006. Eligible children whose parents gave written informed consent to participate in the study were withdrawn from the classrooms for testing by Karina Pont.

Each child was seated at an appropriately sized table and chair with Pont seated opposite. The child was informed about the study and assent to participate was confirmed. The child was then asked to draw a picture of a ‘smiley face,’ which was used to determine the preferred hand. He or she then used the preferred hand to complete the TIHM, which was scored at the time of administration. When specific consent was provided, children’s assessments were videotaped for use in evaluating interrater reliability. The camcorder was positioned to focus on the child’s preferred hand from in front of and to the side opposite the preferred hand.

Test–Retest Reliability. To examine test–retest reliability, each child was tested on two separate occasions 2 weeks apart. We chose this retest period because we believed it to be short enough to preclude hand-skill development and long enough to offset any immediate learning effects. We used 29 completed data sets to examine test–retest reliability.

Interrater Reliability. We selected 27 videotapes on the basis of image quality, inability to identify the child, and availability at the time of distribution to raters, given that distribution occurred prior to finalization of data collection. Three occupational therapists and two occupational therapy students rated videotapes, resulting in 100 paired data sets for the analysis of interrater reliability. The raters participated in a 2-hr training workshop before scoring the videotapes.

Data Analysis

We subjected all data to Rasch analysis using the Winsteps program (Version 3.62.1; Linacre, 2006) according to the partial-credit scale model, allowing for the possibility of different items having different rating scales on the same test. Rasch modeling provides valuable information about a test at the item level and about its rating scale structure. This level of detail is not available using classical test development processes. Quality items were reverse scored so that low scores were most desirable in all cases.

During Rasch analysis, ordinal data are transformed into interval data through a series of logarithmic calculations. All items and children were placed on the same hierarchical scale according to their relative difficulty or ability, respectively, on the basis of the Rasch model’s assumptions (i.e., all children are more likely to receive better scores on easy items than on hard items, and children with greater ability are more likely to score better on all items than are children with less ability; Bond & Fox, 2001). The positions of the items and children on the difficulty–ability continuum are reported as measure scores and expressed as logits (log-odds probability units) or rescaled units (logits + 50).

The TIHM’s construct validity can be verified by the extent to which its items are ordered in a logical hierarchy of item difficulty according to the clinical and theoretical construct of IHM and how well the items fit the expectations of the Rasch model. Goodness-of-fit statistics for each item are reported in the form of mean square (MnSq) and Z standard distribution statistics (Zstd), with ideal values of 1 and 0, respectively, and acceptable values of 0.6–1.4 and –2.0–2.0, respectively (Linacre, 2006). A MnSq > 1.4 and Zstd > 2.0 indicate that the scores for that item are unpredictable or erratic, suggesting that the item is not measuring the same construct as the other items or that the item is
being misinterpreted by those scoring it. *MnSq* values < 0.6 and *Zstd* values < −2.0 indicate overpredictability of the item in the test and may suggest item redundancy (Bond & Fox, 2001; Linacre & Wright, 1994). When both the *MnSq* and *Zstd* values for both infit and outfit are outside the acceptable values, the item is considered to fail to conform to the expectations of the Rasch model. Infit statistics are sensitive to unexpected scores on items at the child’s overall ability level, whereas outfit statistics are more sensitive to unexpected scores on items that are modeled as relatively easy or hard for the child (Bond & Fox, 2001). Conventionally, an instrument is considered to have adequate construct validity when 95% or more of the items fit the expectations of the Rasch model.

Winsteps (Linacre, 2006) provides information explaining the variance that exists in a tool through residual-based PCA. Unlike common factor analysis, PCA with residuals examines variance in a test that cannot be explained by the construct being measured (in this case, the principal component, IHM) through identification of contrasting factors (Linacre, 2006). The magnitude of the residual contrast is measured in eigenvalue units (*EUs*). A test is considered unidimensional when > 60% of the variance in scores can be explained by the Rasch dimension and when the *EU* of the first residual contrast is < 3.0. An *EU* of 3.0 indicates that the residual contrast has a strength of three items. Values > 3.0 suggest that the test may be multidimensional (Linacre, 2006).

Winsteps (Linacre, 2006) also provides information about the way in which participants achieve on test items, referred to as the *rating scale structure*. To provide useful information about a rating scale structure, it is recommended that a scale have no fewer than 10 responses per category. A good rating scale structure should achieve four goals: (1) it must make sense and be logical and practical; (2) it must have a uniformly shaped and unimodal distribution; (3) average measures of each category must increase evenly according to the theoretical difficulty, such that children with higher ability are more likely to achieve better scores than those with lower ability; and (4) each category must represent a distinct proportion of the variable being tested (Bond & Fox, 2001; Linacre, 1999, 2006; Wright & Linacre, 1992).

The test–retest reliability and interrater reliability of a test can also be examined by means of the Rasch model. In Rasch analysis, a measurement of error (standard error [SE]) is given for each measure score, for items and children. An error band is created by adding and subtracting the *SE* from each child’s measure score. The error bands for the first and second testing occasions and for each rater are plotted. When the error bands overlap, the child’s measure of ability is shown to be statistically similar. Adequate test–retest and interrater reliability are demonstrated when 95% or more of the error bands overlap.

We conducted data analysis in three phases. In Phase 1, we examined the original data set for construct validity. During Phase 2, we adapted the scale to improve its validity by removing items identified in Phase 1 as being redundant or not conforming to the expectations of the Rasch model. In Phase 3, we altered the rating scale structure according to Rasch guidelines (see Bond & Fox, 2001; Linacre, 2006) until we identified the most clinically useful tool. We then examined the modified data set for evidence of construct validity, test–retest reliability, and interrater reliability (see Figure 1).

**Results**

A total of 45 typically developing children participated in the study (27 boys, 18 girls). Their mean age was 71 months (*SD* = 3.1, range = 66–78). Five (11.1%) of the participating children were left-hand preferred.
Phase 1: Examination of the Original Data Set

We examined 40 items in the first phase of data analysis: time scores, drop scores, stabilization scores, and quality ratings for each of the five tasks (rotation, translation of two pegs [Trans2], translation of three pegs [Trans3], translation of four pegs [Trans4], and translation of five pegs [Trans5] for two trials each—i.e., four ratings × five tasks × two trials). We examined scores for both trials to obtain the most information about each child and each item. We first examined the evidence for validity of the test.

Goodness of Item Fit and PCA. Examination of the TIHM’s goodness of item fit revealed that data from 17 items (43%) did not conform to the expectations of the Rasch model in that they were too unpredictable (MnSq > 1.4, Zstd > 2). Those items included rotation drop and stabilization scores for both Trial 1 and Trial 2; Trans2 time and drop scores for Trial 2; Trans3 time and drop scores for Trial 2 and stabilization scores for Trials 1 and 2; Trans4 drop scores for Trial 2 and stabilization scores for Trials 1 and 2; and Trans5 drop and stabilization scores for Trials 1 and 2. In addition, data from all the quality items (10 items) failed to conform to the expectations of the Rasch model because the goodness-of-fit scores were below the acceptable range (MnSq < 0.4, Zstd < –2.0), suggesting overpredictability and possible redundancy of all the quality items.

The first residual contrast of the PCA had a magnitude of 4.9 EU, suggesting that the test may not be unidimensional. In this contrast, quality ratings and time scores separated two groups. The contrast identified a group of 10 tests (7%) on which the TIHM tasks were completed quickly at the expense of completing the tasks well. Some children appeared in the contrast more than once as a result of multiple observations of the same child. Another group of 19 tests (13%) were completed well at the expense of speed. This residual contrast identified two seemingly different reactions of the children toward the testing, which may account for the large EU. The two strategies adopted by children completing the TIHM meant that items differed in difficulty for different children, violating a basic assumption of the Rasch model.

Phase 2: Adaptation of the TIHM by Removing Items

During Phase 2, we made two adaptations to the TIHM to improve its construct validity. Quality scores (10 items) were removed from the analysis because they did not conceptually fit the construct. We also removed two tasks (Trans2 and Trans3, 12 items) from the analysis as a result of a pattern that was identified by the PCA. This pattern identified a group of 10 children (7%) whose performance deteriorated as the test progressed. Conversely, the performance of another group of 10 children (7%) improved even though the items increased in difficulty, presumably through better understanding of the task or how to manipulate the pegs more efficiently. Once again, a basic assumption of the Rasch model was violated because items were not equally difficult for all children. The results of these two groups converged between Trans3 and Trans4, with Trans4 and Trans5 items providing a more accurate representation of all the children’s IHM skills in this age group. To eliminate the effect of this pattern and the poor goodness of item fit of Trans2 and Trans3, these tasks were removed from the data analysis. Trans2 and Trans3 were made into practice tasks.

Goodness of Item Fit and PCA. After removal of 22 items from the analysis (all quality items and time, drop, and stabilization scores for Trans2 and Trans3), goodness-of-fit statistics revealed that data from the remaining 18 items conformed to the expectations of the Rasch model, providing evidence that this version of the TIHM was unidimensional. The first residual contrast of the PCA was 2.7 EU, providing further evidence that the revised TIHM was unidimensional.

Rating Scale Structure. The rating scale structure of the data set revealed that all items had disordered response patterns. That is, some children with lower average ability measures achieved higher scores on certain items than children with higher average ability measures. The distribution of scores for some items (rotation time scores for Trial 1, Trans4 time scores for Trial 1, Trans4 time scores for Trial 2, Trans5 time scores for Trial 1) were very uneven.

Phase 3: Adaptation of the TIHM by Altering the Disordered Rating Scale Structure

Although PCA and goodness-of-fit statistics were acceptable in Phase 2, the small numbers of children responding to each category and the disordered response patterns suggested that collapsing the item responses might improve the test’s validity and reliability. That is, by reducing the number of possible scores a child could receive, the children were more likely to receive the same score on another administration, increasing the reliability of the revised TIHM. Moreover, an ordered scale structure is necessary for a valid test. We determined new item response categories by creating a more uniform distribution of respondents in each category and changing the average measures of the adjacent categories, producing a more clinically useful tool (Linacre, 2006). All item responses were collapsed so that a higher measure is more desirable, indicating more ability. As an example, Table 1 compares the average item response measure scores for Trial 2 Trans4 time scores, before and after collapsing the item categories. In Trans4, children who completed the task in 9 s or less received a time score of 4, children who completed the task
Table 1. Average Item Response Measure Scores for Trans4 Time
Scores for Trial 2: Before and After Collapsing Rating Scales

<table>
<thead>
<tr>
<th>Time (seconds)</th>
<th>Frequency</th>
<th>Average Measure</th>
<th>Score</th>
<th>Frequency</th>
<th>Average Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>2</td>
<td>47.02</td>
<td>4</td>
<td>26</td>
<td>78.95</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>47.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>47.88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>47.94</td>
<td>3</td>
<td>58</td>
<td>70.35</td>
</tr>
<tr>
<td>11</td>
<td>41</td>
<td>48.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>48.43</td>
<td>2</td>
<td>50</td>
<td>64.99</td>
</tr>
<tr>
<td>13</td>
<td>20</td>
<td>48.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>48.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>48.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>48.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>48.61</td>
<td>1</td>
<td>9</td>
<td>56.98</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>48.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>48.80</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Trans4 = translation of four pegs.

Table 2. Phase 3: Goodness of Item Fit Statistics

<table>
<thead>
<tr>
<th>Item</th>
<th>Item Measure</th>
<th>SE</th>
<th>Infit</th>
<th>Zstd</th>
<th>Outfit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>MnSq</td>
<td>Zstd</td>
<td>MnSq</td>
</tr>
<tr>
<td>Trans5, stabilizations, Trial 2a</td>
<td>45.4</td>
<td>1.4</td>
<td>1.47</td>
<td>3.3</td>
<td>1.44</td>
</tr>
<tr>
<td>Rotation, stabilizations, Trial 1</td>
<td>43.1</td>
<td>1.5</td>
<td>1.34</td>
<td>2.3</td>
<td>1.38</td>
</tr>
<tr>
<td>Rotation, stabilizations, Trial 2</td>
<td>46.8</td>
<td>1.3</td>
<td>1.30</td>
<td>2.4</td>
<td>1.20</td>
</tr>
<tr>
<td>Trans4, drops, Trial 2</td>
<td>43.4</td>
<td>1.5</td>
<td>1.11</td>
<td>0.9</td>
<td>1.29</td>
</tr>
<tr>
<td>Trans5, stabilizations, Trial 1</td>
<td>41.0</td>
<td>1.6</td>
<td>1.28</td>
<td>1.8</td>
<td>0.99</td>
</tr>
<tr>
<td>Trans4, stabilizations, Trial 1</td>
<td>42.9</td>
<td>1.5</td>
<td>1.11</td>
<td>0.9</td>
<td>1.29</td>
</tr>
<tr>
<td>Trans5, drops, Trial 1</td>
<td>50.5</td>
<td>1.2</td>
<td>1.14</td>
<td>1.3</td>
<td>1.20</td>
</tr>
<tr>
<td>Rotation, drops, Trial 1</td>
<td>34.3</td>
<td>2.0</td>
<td>1.09</td>
<td>0.5</td>
<td>0.86</td>
</tr>
<tr>
<td>Rotation, drops, Trial 2</td>
<td>29.4</td>
<td>2.5</td>
<td>1.06</td>
<td>0.3</td>
<td>0.97</td>
</tr>
<tr>
<td>Trans5, time, Trial 1</td>
<td>67.0</td>
<td>1.2</td>
<td>0.87</td>
<td>−1.1</td>
<td>1.04</td>
</tr>
<tr>
<td>Trans4, stabilizations, Trial 2</td>
<td>39.0</td>
<td>1.7</td>
<td>1.03</td>
<td>0.3</td>
<td>0.82</td>
</tr>
<tr>
<td>Trans5, drops, Trial 2</td>
<td>56.0</td>
<td>1.2</td>
<td>1.03</td>
<td>0.4</td>
<td>1.00</td>
</tr>
<tr>
<td>Trans5, time, Trial 2</td>
<td>68.2</td>
<td>1.2</td>
<td>0.91</td>
<td>−0.7</td>
<td>0.97</td>
</tr>
<tr>
<td>Trans4, time, Trial 1</td>
<td>65.3</td>
<td>1.2</td>
<td>0.86</td>
<td>−1.3</td>
<td>0.87</td>
</tr>
<tr>
<td>Trans4, time, Trial 2</td>
<td>62.3</td>
<td>1.2</td>
<td>0.81</td>
<td>−1.9</td>
<td>0.83</td>
</tr>
<tr>
<td>Rotation, time, Trial 2</td>
<td>61.3</td>
<td>1.1</td>
<td>0.78</td>
<td>−2.4</td>
<td>0.78</td>
</tr>
<tr>
<td>Rotation, time, Trial 1</td>
<td>64.2</td>
<td>1.2</td>
<td>0.72</td>
<td>−2.9</td>
<td>0.74</td>
</tr>
<tr>
<td>Trans4, drops, Trial 1</td>
<td>40.0</td>
<td>1.6</td>
<td>0.72</td>
<td>−1.9</td>
<td>0.58</td>
</tr>
<tr>
<td>M</td>
<td>50</td>
<td></td>
<td>1.04</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>SD</td>
<td>11.8</td>
<td></td>
<td>0.22</td>
<td>1.8</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Note. Trans5 = translation of five items; Trans4 = translation of four items.

Table 2. Phase 3: Goodness of Item Fit Statistics

in 10 s to 11 s were given a score of 3, children who completed the task in 12 s to 16 s received a score of 2, and children who completed the task in more than 17 s were given a score of 1. All items were collapsed using similar principles. After the item rating scale structure was collapsed, the children were receiving a greater range of scores on the revised TIHM’s hierarchical scale of ability–difficulty (49.29–100.05 rescaled units) and items (29.39–68.23 rescaled units).

We then examined the collapsed data set, with quality ratings for Trans2 and Trans3 tasks removed, for evidence of construct validity, test–retest reliability, and interrater reliability.

Goodness of Item Fit. One of the 18 items (5.6%), Trans5 stabilization Trial 2, failed to conform to the expectations of the Rasch model. Table 2 shows the MnSq and Zstd values, measures of item difficulty, and SE for each item.

PCA. The PCA provided additional evidence of the unidimensionality of the revised test. The variance in the TIHM explained by the principal component (IHM) was 72.6% (modeled variance = 72.1%), with the first residual contrast having a magnitude of 2.6 EU. Table 3 contains the final test items and scoring criteria of the most useful version of the TIHM, the Test of In-Hand Manipulation—Revised Version (TIHM–R).

Test–Retest Reliability. We plotted the error bands of ability measure (measure score ±1 SE) for each child for the first and second TIHM–R testing occasions. Error bands for 22 of the 29 ability measures (75.86%) overlapped. Because this overlap was substantially lower than the desired 95%, adequate test–retest reliability of the TIHM–R was not supported. We observed three patterns of performance when
Table 3. Test of In-Hand Manipulation–Revised Items and Scoring Criteria

<table>
<thead>
<tr>
<th>Task</th>
<th>Item</th>
<th>Scoring Criteria</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rotation task</td>
<td>Time</td>
<td>&lt;10 s</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11–12 s</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13–15 s</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;16 s</td>
<td>1</td>
</tr>
<tr>
<td>Trans2 task (practice task)</td>
<td></td>
<td>No score</td>
<td></td>
</tr>
<tr>
<td>Trans3 task (practice task)</td>
<td></td>
<td>No score</td>
<td></td>
</tr>
<tr>
<td>Trans4 task</td>
<td>Time</td>
<td>&lt;9 s</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10–11 s</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12–16 s</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;17 s</td>
<td>1</td>
</tr>
<tr>
<td>Trans5 task</td>
<td>Time</td>
<td>&lt;10 s</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11–15 s</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16–19 s</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;20 s</td>
<td>1</td>
</tr>
<tr>
<td>Rotation, Trans4, and Trans5 tasks</td>
<td>Drops and stabilizations</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;2</td>
<td>2</td>
</tr>
</tbody>
</table>

Note. Trans1 = translation of one item; Trans2 = translation of two items; Trans3 = translation of three items; Trans4 = translation of four items; Trans5 = translation of five items.

Examining the TIHM–R’s test–retest reliability. The first observed pattern was that children who had more errors (drops or stabilizations) during the TIHM–R tasks were very likely to have dissimilar scores between Test 1 and Test 2. The second pattern was that the second test scores of children with dissimilar retest scores tended to regress toward the mean; some children’s TIHM–R scores improved, and others had lower ability measures in their second test. The final pattern observed during the analysis of the TIHM–R’s test–retest reliability was that differences in the performance of translation with stabilization tasks leading to dissimilar retest results were more common than differences in rotation tasks.

Interrater Reliability. We plotted the error bands of ability measure awarded by each rater to determine whether any differences in scores were likely to be the result of differences in the severity of the raters. The measure bands of all children overlapped when scored by two or more raters. Forty-six of the 100 data sets were given exactly the same overall ability measure by two or more raters. Very high interrater reliability of the TIHM–R was established.

Discussion

In this study, we used Rasch modeling to examine the construct validity and aspects of reliability of Case-Smith’s (1996) TIHM. We found limited evidence for validity of the original version of the TIHM because of the large number of items (43%) that did not conform to the expectations of the Rasch model. Using the data obtained, we revised the test. In particular, it was clear that the quality ratings (which were added after the original studies; e.g., Case-Smith, 1995) did not conform, and these were removed from the test. In addition, translation with fewer than four pegs did not accurately represent the IHM skill of this population, and thus translation with two and three pegs were converted into practice tasks. This configuration of test items, however, may not be appropriate for younger children and needs to be tested in a wider range of age groups.

Collapsing the TIHM’s rating scale structure increased its clinical utility. A reduction in the number of possible scores and having the same number of scoring categories for each item enables raters to score the test with greater ease. Moreover, collapsing the rating scale structure increased the test’s sensitivity in this age group. The goodness of item fit statistics for the revised TIHM with the collapsed rating scale revealed that all but one item (94.4%) conformed to the expectations of the Rasch model, which we deemed acceptable. This item, Trans4 stabilization for Trial 2, remained part of the revised test.

The resultant TIHM–R is an 18-item test involving rotating pegs and translating four and five pegs into and out of the palm. Translations of two and three pegs are included as practice items. Two trials are given, during which time and numbers of drops and stabilizations for each item are recorded on a scale ranging from 1 to 4 (see Table 3).

Interrater reliability of this version of the TIHM–R was very high, indicating that the TIHM–R scores are resistant to variation resulting from differences in raters’ scoring. Interrater reliability may have been enhanced by raters’ attendance at a 2-hr training workshop. Examination of interrater reliability with untrained raters should be the topic of future research because it more closely conforms to use by practitioners.

The TIHM–R’s inadequate test–retest reliability indicates susceptibility of test scores to change over time across a 2-week retest period in typically developing children in this age group. While examining the TIHM–R’s test–retest reliability, we observed that children who dropped or stabilized the pegs more frequently than their peers had lower test stability over a 2-week retest period, perhaps because IHM is still developing until age 7 (Exner, 1992; Pehoski et al., 1997a, 1997b). Some children may compensate for underdeveloped IHM by stabilizing objects on an external surface. Alternatively, they may be prone to dropping objects. The lack of test–retest stability implies that the TIHM–R cannot be confidently used as an outcome measure in this age group.

Two factors may limit the usefulness of the TIHM–R. The first is that the TIHM–R assesses only complex rotation and translation with stabilization; shift and simple rotation are not included. Shift and simple rotation may be important to the performance of some fine motor tasks; however, these relationships cannot be examined using the TIHM–R. The
second factor is that the TIHM–R may not be sensitive to possible differences in the performance of finger-to-palm translation and palm-to-finger translation, as these are assessed in one task. Several children participating in this study were observed to have less ability with palm-to-finger translation than with finger-to-palm translation. The difficulties these children experienced with palm-to-finger translation were often masked by the child’s adequate finger-to-palm translation skills.

Limitations

The small sample size of 45 children (29 for test–retest reliability) and the inclusion criteria of typically developing children between 5.5 and 6.5 years old limits the generalizability of the study’s results to other age groups or children with suspected dysfunction.

Conclusions and Implications for Future Research

IHM is considered an essential component of the fine motor skills necessary to manipulate tools and objects and to master the world. Specific IHM activities are frequently included in occupational therapy intervention sessions for children with fine motor difficulties (Case-Smith, 2000). The TIHM–R is an assessment of IHM that examines rotation and translation with stabilization. No previous study has explored the TIHM’s psychometric properties. We used Rasch modeling to create a version of the test that has evidence of adequate construct validity and excellent interrater reliability when used with typically developing children between the ages of 5.5 and 6.5 years. The TIHM–R may be useful for examining relationships between IHM and other aspects of fine motor dysfunction. The TIHM–R may also be useful for identifying children with IHM difficulties once normative data are developed; however, its inadequate test–retest reliability over a 2-week period limits its use as an outcome measure. The TIHM–R demonstrates potential to be a useful assessment of IHM. Before the TIHM–R can be confidently used in occupational therapy practice, it needs to be validated with a wider age range and with children of different levels of ability and diagnoses. ▲

Acknowledgments

This study was completed by Karina Pont as part of the requirements for the completion of the Honours Program for Bachelor of Applied Science (Occupational Therapy), under the supervision of Margaret Wallen and Anita Bundy. We thank the children and staff of participating schools for their participation; Heidi Willis for assistance in data collection; Betina Frazia, Katrina Beissel, and Heidi Willis for participation in the study; and J. Michael Linacre for assistance with data analysis and critique of the manuscript. This study was supported in part by a grant from the School of Occupation and Leisure Sciences in the Faculty of Health Sciences at the University of Sydney.

Note

Three children were not present at the time of the second testing and were excluded from the test–retest portion of the study. Thirteen other children (from one school) were not included in the test–retest portion of the study because a testing environment that complied with TIHM administration guidelines was not available at the second testing occasion.

References


Exner, C. E. (1992). In-hand manipulation skills. In J. Case-Smith & C. Pehoski (Eds.), Development of hand skills in the
child (pp. 35–45). Rockville, MD: American Occupational
Therapy Association.
Scales. Austin, TX: Pro-Ed.
of in-hand manipulation and relationship with activities.
Jewell, K., & Humphry, R. (1993). Reliability of an observa-
tion protocol on in-hand manipulation and functional skill
development. Physical and Occupational Therapy in Pediatrics,
13(3), 67–81.
[Computer software and manual]. Retrieved August 6, 2006,
from www.winsteps.com
Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit
values. Rasch Measurement Transactions, 8(3), 370. Retrieved
October 3, 2006, from www.rasch.org/rmt/rmt83b.htm
science research: A handbook of quantitative methods. Sydney,
New South Wales, Australia: Allen & Unwin.
Pehoski, C., Henderson, A., & Tickle-Degnen, L. (1997a). In-
hand manipulation in young children: Rotation of an object
in the fingers. American Journal of Occupational Therapy, 51,
544–552.
Pehoski, C., Henderson, A., & Tickle-Degnen, L. (1997b). In-
hand manipulation in young children: Translation move-
ments. American Journal of Occupational Therapy, 51,
719–728.
positioning quality on typical 6- and 7-year-old children’s
object manipulation skills. American Journal of Occupational
Therapy, 58, 380–388.
Wright, B. D., & Linacre, J. M. (1992). Combining and split-
Ziviani, J., & Wallen, M. (2006). The development of grapho-
motor skills. In A. Henderson & C. Pehoski (Eds.), Hand