Comparison of the Test–Retest Reliability of the Work Box™ Using Three Administrative Methods

Loucretia Speller, Judith A. Trollinger, Patti A. Maurer, Craig E. Nelson, David F. Bauer

Key Words: industrial therapy • work evaluation

Objective. The purpose of this study was to compare the test–retest reliability of three administrative methods of the Work Box™: (a) the original instructions, (b) a revised version of the original instructions, and (c) another revised version that was based on suggestions made by authors of the first two versions of the instructions.

Method. Sixty subjects without disabilities were randomly grouped so that 20 subjects were tested per administrative method. The assessment was administered to each subject on two occasions, with a 7-day to 14-day period between tests. Scores were recorded as time in seconds, and intraclass correlation coefficients (ICCs) were used to calculate the reliability.

Results. The ICCs for assembly, disassembly, and total scores were .589, .604, and .654, respectively, for the original instructions; .424, .572, and .545 for the revised instructions; and .781, .579, and .717 for the second revised instructions. Reliability was found to be higher for men than for women and for subjects who claimed to have more rather than less experience in similar manual dexterity tasks.

Conclusions. On the basis of the reliability of each administrative method and comments made by subjects about their understanding of the instructions, the second revised version of the instructions is recommended as the standard method. The results also indicate that the assessment is most appropriate for a population of men with manual dexterity experience. With further standardization, the Work Box could be a valuable assessment tool for therapists working in industrial rehabilitation settings.

Growth of private-sector industrial rehabilitation, changes in workers’ compensation laws, and increasing costs of industrial rehabilitation have created a need for occupational therapists to specialize in this area of practice (Matheson, Ogden, Violette, & Schultz, 1985; Niemeyer, Jacobs, Reynolds-Lynch, Bentoncourt, & Lang, 1994). An additional factor is the notable problem of cumulative trauma disorders (Louis, 1995). Occupational therapists involved in industrial rehabilitation provide work-oriented treatment to facilitate the injured worker’s transition from patient to productive worker (Holmes, 1985; Jacobs, 1995; Schultz-Johnson, 1991; Wtyick, Niemeyer, Ellexson, Jacobs, & Taylor, 1991). To facilitate the transition from patient to worker, a thorough evaluation of the patient is needed (Jacobs, 1995) by a therapist who understands the true value of each assessment tool (Veloza, 1993). A work tolerance assessment is one general tool that occupational therapists use to determine the extent to which a patient is able to perform a specific work-related activity. Matheson (1986) divided the various work tolerance screening
assessments into four categories on the basis of their design:

1. Psychometric tests of motor speed and dexterity that measure manipulation skills with various objects
2. Demand-structured work samples, such as the Work Box\(^\text{TM}\), that are “highly structured simulated work tasks” involving materials and tools that are identical to those in an actual job\(^*\) (p. VI-6)
3. Work capacity evaluations that simulate specific work tasks and can be adjusted to increase the demand of the tasks
4. Structured work simulations, which are workstations at the evaluation site that realistically replicate actual jobs

Most occupational therapists working in industrial rehabilitation use a variety of screening tools from all four categories to obtain an accurate picture of a patient’s level of function, which can then influence the treatment plan. Efforts to develop and refine work evaluations continue (Bhambhani, Esmail, & Brintnell, 1994; Cetnik, Renfro, & Coleman, 1995; Esmail, Bhambhani, & Brintnell, 1995; McClure & Flowers, 1992; Mooney & Matheson, 1994).

One desirable feature of work samples is test–retest reliability (i.e., the stability of a test over time) (Johnston, Findley, DeLuca, & Katz, 1991). To determine test–retest reliability, the same test is administered to the same group at two different times, and the closer the test scores are, the more reliable the test (Walsh, 1989).

Numerous factors influence test–retest reliability. Variations between trials in the test conditions, such as change in location, difference in environmental temperature, or variations in test setup, can affect test results (Bolton, 1974; Cunningham, 1986). Reliability of a scorer also influences test–retest reliability (Cunningham, 1986). If there is only one scorer, this factor is not much of an issue as long as the scorer is consistent with each subject and between trials. When there is more than one scorer, it is important that pilot testing be done to ensure that all scorers are consistent. Finally, group homogeneity affects the reliability of tests (Deitz, 1989; Graham & Lilly, 1984; Mehrens & Lehmann, 1991). If a group has very homogenous scores, the reliability coefficients will most likely be low, resulting in an overall low test–retest reliability.

Because the same test is given to a person twice for test–retest reliability, a practice effect can potentially influence results. The practice effect is a significant difference in score from the first trial to the second as a result of the person learning from the first trial. A practice effect has been found in evaluations involving manual dexterity skills (Black, Nelson, Maurer, & Bauer, 1993; Stern, 1992; Tiffin & Asher, 1948) due, in part, to the motor learning that occurs with practice, which eventually improves the skill (Kortke, 1980). Karger and Hancock (1982) identified several task and human factors that affect the rate at which a person learns repetitive jobs and that apply to the learning of manual dexterity skills used in work tolerance evaluations. Among these factors are task cycle length (rate slows with longer cycle), position and movement complexity (rate slows for more complex tasks), prior related training (rate slows when there has been no related training), age (rate slows after years of peak learning), and participation in learning new tasks (rate slows with aging without ongoing exposure to new learning). Anastasi (1988) and Arkava and Snow (1978) noted that the longer the time interval between trials, the less influence the first trial will have on the second because there is a greater opportunity to forget learned skills. All these factors can influence test–retest reliability.

The Work Box

The Work Box\(^1\) (formerly known as the Assembly Box) is an assessment used to evaluate the manual dexterity and work skills of persons who have sustained injuries (e.g., hand, back) (Black, 1991). It incorporates several worker skills, including lifting, moving, lowering, aligning holes, assembling and disassembling wing nuts and bolts, and problem solving. Methods–time measurement analysis divides a manual operation into the basic motion elements required to perform the task in a predetermined standard time (Swanson, Swanson, & Goran-Hagert, 1990). Black (1991) used this analysis to determine that the Work Box assesses manual dexterity, nonmanipulative upper-extremity motion, and assembly and disassembly activities used in various work sites. The Work Box consists of two square end plates, each with a side dimension of 40.5 cm (16 in.). To these end plates, 16 crossbars measuring 62.5 cm by 2.5 cm (24.5 in. x 1 in.) are attached via nut-and-bolt assembly (see Figure 1).

Development of Instructions

Black (1991) developed the first version of standard instructions for administration, which require the advanced attachment of one of the crossbars by the evaluator to hold the end plates in position. The instructions also specify the placement of the materials and the order in which the crossbars are to be attached. Following assembly of three crossbars to reduce a practice effect and their

---

\(^1\) Available from Advanced Therapy Products, Inc., PO Box 3420, Glen Allen, Virginia 23058-3420.
subsequent removal, the subject is timed on the assembly of 15 crossbars, using wing nuts and bolts to manually secure each bar to the end plates. After assembling the top side, the box is rotated away from the subject so that the next side (now on top) can be assembled. This procedure is repeated until all sides are assembled. The assembly time is recorded in seconds. After a 10-min rest period, the subject is allowed to practice disassembling four crossbars by unfastening the wing nut and bolt at each end of each bar. These crossbars are reattached by the evaluator, and the subject is timed while disassembling all 15 crossbars. Disassembly time is recorded in seconds. The evaluator records comments made by the subject during the test, the number of wing nuts and bolts dropped, and other observations. The total test score is calculated by adding the assembly and disassembly times.

**Previous Research**

Black et al. (1993) studied the test-retest reliability of the Work Box for 28 subjects receiving industrial rehabilitation services. The ICCs for the group were .720 for assembly, .740 for disassembly, and .746 for the total score. On the basis of standards set by Fleiss (1986), these values were just below the level for excellent reliability. Black et al. found higher reliability of the men's scores versus the women's scores, which may have been due to the small female sample size. Other limitations to the study included random threading flaws in the wing nuts and bolts, failure to specify the tightness of the wing nuts, and the short time between test and retest.

Feden (1993) addressed these limitations by modifying the administration instructions. He included that the wing nuts are to be tightened one-quarter to one-half turn and that the scorer checks the threading of all wing nuts and bolts before each test. Feden deleted the final rotation of the box during assembly. The time interval between trials was also lengthened to between 7 days and 14 days. Using 31 female college students as subjects, Feden determined the ICCs to be .370 for assembly, .826 for disassembly, and .653 for the total score. Although the assembly and total scores were lower than those in Black et al. (1993), there was less of a mean change from the test to the retest and a smaller standard error of measurement (SEM). The fact that Feden used a homogenous sample of women without disabilities may have caused the reliability to be underestimated. Additionally, none of Feden's subjects reported being previously involved in activities that involved nut-and-bolt assembly. Because of the discrepancies between the results of Feden and Black et al. and the need to further revise the administration methods, the present study attempted to clarify the administrative methods and to determine test-retest reliability with another subject group. The specific research question, therefore, was: What is the test-retest reliability for three administrative methods of the following scores on the Work Box: assembly, disassembly, and total?

**Method**

**Subjects**

A convenience sample of 60 subjects was drawn from three pools: students from a university in Richmond, Virginia (n = 42), auto mechanics students from a Richmond-area community college (n = 10), and contract archaeologists (n = 8). All persons who volunteered to be subjects were tested as long as they met the following inclusion criteria: (a) between the ages of 14 and 60 years, (b) no medical history that could interfere with performance, and (c) no previous experience with the Work Box. Of the 60 subjects, 21 were men, and 39 were women. Their ages ranged from 18 to 57 years (M = 28 years). Fifty subjects identified themselves as right-hand dominant. Thirty-one claimed that they had no previous experience with tasks involving manual dexterity skills similar to those used with the Work Box (see Table 1). Consent forms were completed before testing.

**Procedure**

We revised the administration instructions for the Work Box on the basis of suggestions by Black (1991) and Feden (1993). The tightness of the wing nuts and bolts was changed from one-quarter to one-half turn (as in Feden's instructions) to exactly one-quarter turn in order to make the tightness more standard. After the practice period and before the timed assembly test, the data collector squared the Work Box by using one crossbar with two wing nuts to realign the end plates and supporting horizontal bar. Squaring the Work Box before testing was done because of the frustration noted by subjects in Feden's study who had difficulty attaching the bars due to misalignment.
Table 1
Demographic Variables for Subjects Tested With the Work Box™ as a Function of Administration Method

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Female</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Hand dominance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td>17</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>Left</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Experience</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>11</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>No</td>
<td>9</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Age range</td>
<td>20-46</td>
<td>18-47</td>
<td>18-57</td>
</tr>
<tr>
<td>Mean age</td>
<td>28.2</td>
<td>26.9</td>
<td>28.9</td>
</tr>
</tbody>
</table>

Note. N = 60.

In a pilot test, the two data collectors practiced administering each of the three methods (i.e., Black's, Feden's, revised) to six persons. On the basis of the pilot testing results, final revisions were made to the revised instructions to specify tightening each wing nut one-quarter turn after making contact with the end plate.

Subjects were tested at sites from where they were recruited. The test site for the university student sample was one of five similar classrooms, with subjects facing the back of the classroom and no extraneous noises present during most test sessions. The second test site was a workbench in the back of an auto mechanics classroom. Testing for this sample occurred during class time, often with loud background noises and other distractions. The test site for the contract archaeologist sample was a local historical site at a table adjoining the storage and maintenance areas; there was a constant background hum of machinery and occasional distractions.

Subjects were assigned to one of three administrative methods by randomization in blocks of six at each site. By randomizing with a block size of six, the number of subjects receiving each of the three administration methods was equalized at every sixth subject. This blocking was used to ensure an equal distribution of administrative methods among the samples.

On the initial test day, the subjects completed a demographic questionnaire, the data collector read them the assembly instructions, and they practiced assembling three crossbars. Subjects' performance was then observed and timed in seconds. A 10-min or 5-min rest break, depending on the administration method, was given before disassembly. During this break, a brief interview was conducted to determine factors affecting performance, including experience with tasks that involved manual dexterity. The nut-and-bolt assemblies were also checked by the data collector for tightness and corrected as needed for disassembly when Feden's and the revised administrative methods were used. After the break, the data collector read the disassembly instructions to the subject, the subject practiced disassembling three crossbars, and performance was observed and timed in the same manner as for assembly.

Subjects were retested between 7 days and 14 days after the initial test and at the same time of day as the initial test. The same administrative method, data collection, and testing procedures were used for both trials. Subjects were not informed of their results.

Results

Means and standard deviations for assembly, disassembly, and total scores were calculated for each administration method (see Table 2). The mean change between test and retest was also computed. Using ICCs, the test-retest reliability for assembly, disassembly, and total test scores was determined, along with the 95% upper and lower confidence bounds for each ICC. On the basis of these confidence intervals, no administration procedure was determined to be significantly better than another. The SEM was also computed.

ICCs and associated confidence bounds were computed for different gender, hand dominance, test site, and experience with similar hand dexterity tasks in order to determine whether these factors affected the reliability. Of these factors, experience (see Table 3) and gender (see Table 4) significantly affected reliability as reflected by nonoverlapping confidence intervals.

Discussion

Administration Method

On the basis of the standards set by Munro and Page (1993), the ICCs from using Black's (1991) instructions (assembly = .589, disassembly = .604, total score = .654) showed moderate reliability. The ICC for Feden's (1993) assembly instructions (.424) constituted a low reliability, but the ICCs for disassembly (.572) and total (.545) scores had moderate reliability. From the revised instructions, the ICCs (assembly = .781, disassembly = .579,
total score = .717) were moderately reliable.

Although not significantly different, the reliability coefficients for the revised instructions were higher than those of Black or Feden for assembly, disassembly, and total score. Only the reliability coefficient for disassembly using Black’s instructions was higher than the coefficients for the other two. Subjects reported negative aspects of Black’s administration method. They did not understand why the box needed to be turned a final time during assembly, commented that the rest period between assembly and disassembly was too long, and were unclear as to how tight the wing nuts should be tightened. The only major complaint from subjects about Feden’s administration method was that they were unclear about the tightness of the wing nuts because the instructions were too specific. The tightness of the wing nuts was also a concern of some subjects who were tested with the revised instructions. On the basis of the complaints and the differences in reliability coefficients, the revised administration method is recommended as the standard instructions for the Work Box.

Experience and Gender

The significant difference in the disassembly scores between men and women is in accordance with Black et al.’s (1993) results. On the basis of the standards set by Munro and Page (1993), the assembly ICC for women (.223) had little, if any, strength. Additionally, their disassembly (.416) and total (.321) scores have low reliability. On the other hand, male subjects’ assembly (.915) and total (.940) scores had very high reliability, and their disassembly (.899) score was just below that mark of having high reliability. When comparing subjects’ experience with manual dexterity tasks similar to those performed with the Work Box, significant differences were found. On the basis of the standards set by Munro and Page (1993), the reliability coefficients for subjects with no similar experience had little, if any, strength in assembly (.289) and were low for disassembly (.488) and total (.412) scores. In contrast, for subjects who claimed to have experience with manual tasks similar to those of the Work Box, assembly (.812), disassembly (.732), and total (.820) scores all indicated

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Descriptive Statistics and Test–Retest Reliability for Work Box Scores as a Function of Administration Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>Test Time (sec)</td>
</tr>
<tr>
<td></td>
<td>M</td>
</tr>
<tr>
<td>Black’s (1991) instructions (n = 20)</td>
<td></td>
</tr>
<tr>
<td>Assembly</td>
<td>513.0</td>
</tr>
<tr>
<td>Disassembly</td>
<td>289.6</td>
</tr>
<tr>
<td>Total score</td>
<td>802.6</td>
</tr>
<tr>
<td>Feden’s (1993) instructions (n = 20)</td>
<td></td>
</tr>
<tr>
<td>Assembly</td>
<td>535.2</td>
</tr>
<tr>
<td>Disassembly</td>
<td>307.5</td>
</tr>
<tr>
<td>Total score</td>
<td>842.7</td>
</tr>
<tr>
<td>Revised instructions (n = 20)</td>
<td></td>
</tr>
<tr>
<td>Assembly</td>
<td>535.2</td>
</tr>
<tr>
<td>Disassembly</td>
<td>304.9</td>
</tr>
<tr>
<td>Total score</td>
<td>840.1</td>
</tr>
</tbody>
</table>

Note. ICC = intraclass correlation coefficient; 95% LB = the 95% lower confidence bound of the ICC; 95% UB = the 95% upper confidence bound of the ICC; SEM = standard error of measurement.

*Decrease in seconds from test to retest.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Descriptive Statistics and Test–Retest Reliability for Work Box Scores as a Function of Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>Test Time (sec)</td>
</tr>
<tr>
<td></td>
<td>M</td>
</tr>
<tr>
<td>No previous experience (n = 31)</td>
<td></td>
</tr>
<tr>
<td>Assembly</td>
<td>529.5</td>
</tr>
<tr>
<td>Disassembly</td>
<td>315.8</td>
</tr>
<tr>
<td>Total score</td>
<td>845.3</td>
</tr>
<tr>
<td>Previous experience (n = 29)</td>
<td></td>
</tr>
<tr>
<td>Assembly</td>
<td>535.9</td>
</tr>
<tr>
<td>Disassembly</td>
<td>298.5</td>
</tr>
<tr>
<td>Total score</td>
<td>810.6</td>
</tr>
</tbody>
</table>

Note. ICC = intraclass correlation coefficient; 95% LB = the 95% lower confidence bound of the ICC; 95% UB = the 95% upper confidence bound of the ICC; SEM = standard error of measurement.

*Decrease in seconds from test to retest.

*Significant at p < .05.
good reliability. These findings indicate that the Work Box is an appropriate assessment for persons with a history of performing manual dexterity tasks similar to those used on the tool but of questionable value for evaluating persons without this history or experience.

Generally speaking, after a person is familiar with a skill, practice improves his or her ability to perform the skill until it can be performed in the least amount of time. Additionally, the more complex the skill, the longer it takes to learn and perfect (Karger & Hancock, 1982). Because the assembly part of the Work Box assessment consists of more complex tasks that take a longer period to learn and perfect than the disassembly part, better reliability would be expected for the disassembly task. The reliabilities were higher for disassembly times when assembly reliabilities were low (i.e., Black's and Feden's instructions, no experience, women) but were lower when the assembly reliabilities were high (i.e., revised instructions, experience, men).

The remarkable difference in scores between men and women is not related to a small number of female subjects, as was the case for Black et al. (1993). The number of subjects in our study was 21 men and 39 women compared with Black et al.'s 22 men and 6 women. A more probable reason for the good reliability of male subjects' scores could be related to experience. Seventeen (81%) men claimed to have some type of experience with skills similar to those used with the Work Box (e.g., repairing household items, fixing cars) compared with 12 (31%) women. Therefore, this study confirms the idea that persons skilled at one task will learn a similar task more easily and more quickly than those who have had no experience with such tasks (Karger & Hancock, 1982). With the experience acquired during the practice phases of the Work Box evaluation, there is less room for change in scores for men than for women because the men learned the task faster, making their scores more reliable. This would also explain the low reliability in Feden's (1993) study because he used an all female sample with no reported experience in nut-and-bolt assembly.

**Limitations**

As with all studies, there were some sources of error. The testing environment could have been better controlled by testing every subject at the same site, thus alleviating distractions that may have affected subjects in one test site but not another. Although pilot testing was performed to ensure that the two data collectors were using the same scoring and observation methods, minor individual differences between them may have affected the results (e.g., how they dealt with unexpected distractions).

**Conclusion**

This study of test–retest reliability of the Work Box is one component of standardizing this assessment of manual dexterity. Of the three administration methods used, the revised instructions described in this study were found to be the most reliable, considering differences in scores and comments from subjects. Experience and gender also were found to affect the reliability of the test. Those subjects with previous experience with manual dexterity skills similar to those used with the Work Box performed better than those with no experience. Additionally, men tended to perform substantially better on the test than did women. The good reliability scores for the subjects with similar manual dexterity experience also show that the tool is targeted to the correct population of persons who use such skills. Development of the Work Box depends on further research, including collection of normative data (to further standardize the assessment), studies determining validity, and examination of how experience relates to learning. ▲

**Acknowledgments**

We thank Peggy Alexander for assistance with data analysis, Tammy Staelker for assistance with manuscript preparation, and David Smith of Advanced Therapy Products, Inc., for supplying equipment. This study was completed by the first two authors in partial fulfillment of...
the requirements for the master of science degree, Department of Occupational Therapy, Virginia Commonwealth University, Richmond, Virginia.

References


