Test–Retest Reliability of the Evaluation Tool of Children's Handwriting-Manuscript

Susan M. Diekema, Jean Deitz, Susan J. Amundson

Key Words: reliability of tests • school-based occupational therapy

Objective. This study examined the test–retest reliability of the legibility portion of the Evaluation Tool of Children's Handwriting-Manuscript (ETCH-M).

Method. The sample consisted of 31 first-grade and second-grade students with handwriting dysfunction. The ETCH-M was administered two times, 1 week apart, to participants. The primary investigator acted as sole rater and followed standard scoring procedures.

Results. The reliability coefficients were .63 for total numeral legibility, .77 for total letter legibility, and .71 for total word legibility. Individual task reliability coefficients were generally lower and ranged from .20 (near-point copy) to .76 (alphabet uppercase).

Conclusion. Total letter, total word, and uppercase letter legibility were more stable than total numeral legibility scores and other individual tasks scores. When evaluating handwriting for a child, it is important to consider ETCH legibility scores as only one aspect of a comprehensive evaluation.

The school-based occupational therapy evaluation focuses on a student's performance in the school environment. Typical areas of school-based evaluation and intervention are (a) mobility and transitions, (b) handling of classroom materials, (c) functional written communication, (d) activities of daily living, (e) school routines, and (f) socialization. A problem with handwriting is one of the most common reasons for an occupational therapy referral in school-age children (Oliver, 1990; Reisman, 1991). Handwriting is important for the child in the school setting for many reasons: The amount of time devoted to paper-and-pencil tasks is high (McHale & Cermak, 1992); reduced handwriting skills may lead to an inability to keep up with written work and in taking notes during class time (Graham & Miller, 1980); poor handwriting may affect persistence, motivation, and sense of efficacy (MacArthur & Graham, 1987); and reduced legibility of a student's work may result in lower grades (Briggs, 1970, 1980).

Handwriting problems may originate from difficulties in cognitive, psychosocial, sensorimotor, fine motor, or auditory or visual processing skills. A full evaluation should include discussion with the teacher regarding concerns and observations; review of the student's educational cumulative file; classroom observation of the student performing school activities, specifically a writing assignment; and formal and informal assessments (Benbow, Hanft, & Marsh, 1992). The school-based therapist depends on available standardized assessments and indi-
vidualized classroom observations to evaluate children, provide classroom consultation, and develop intervention programs. Standardized assessments often are used in determining eligibility for children to receive services, documenting change, and serving as a communication tool among members of the educational team (Bonder, 1989).

Several handwriting assessments exist. Each varies greatly in its approach to the type of writing task, writing style (manuscript or cursive), scoring criteria, and age level. Handwriting legibility can be measured on a global level for readability by comparing the child's writing samples to samples in the assessment manual or on a more specific level for the individual components of handwriting (e.g., slant, spacing, letter size). There is great variability in the definition of legibility and the approach used to measure it (Graham, 1986). One definition of legibility is the ability to recognize a letter, number, or word easily and correctly outside of the context of the word, sentence, or phrase (Hasbrouck, Tindal, & Parker, 1994). Because each practitioner evaluating legibility of handwriting brings a different historical experience regarding letter, word, and numeral recognition, determining legibility is a subjective process.

Because handwriting is an important student occupation, Amundson (1995) developed the Evaluation Tool of Children's Handwriting (ETCH) to fill the need for a comprehensive protocol for evaluating handwriting and the challenges inherent in measuring legibility. This handwriting test examines manuscript and cursive handwriting legibility and speed—the foundations for functional written communication. Well-defined standards are used to determine legibility and provide the school-based therapist with useful information for program planning as well as documenting changes in performance over time. The ETCH contains two separate segments for handwriting tasks: one for manuscript (ETCH-M) and one for cursive.

The ETCH-M is the focus of the current study. An interrater reliability study for this measure has been completed by the test developer, and results are included in the examiner's manual. An understanding of the test–retest reliability is also important so that the therapist can evaluate the expected stability of test scores over time when there is no intervention. This understanding facilitates score interpretation by the therapist. Thus, the purpose of this study was to examine the stability of the ETCH-M legibility scores. This was accomplished by using the ETCH-M to test and retest a sample of first-grade and second-grade students in order to address the following research questions:

1. What are the test–retest reliabilities for measures of letter and numeral legibility of the six individual handwriting tasks of the ETCH-M?
2. What is the test–retest reliability of the total letter legibility score (combination of Tasks Ia, Ib, III, IV, V, VI)?
3. What is the test–retest reliability of the total numeral legibility score (combination of Tasks II and V)?
4. What is the test–retest reliability of the total word legibility score (combination of Tasks III, IV, V, VI)?
5. What are the magnitudes of difference between test and retest scores for each of the following ETCH scores: total letter legibility, total numeral legibility, total word legibility, and the six individual writing tasks?

Method

Participants

A convenience sample of 31 children (24 boys, 7 girls) with identified handwriting performance deficits was tested. The sample was taken from seven elementary schools in the Seattle Public School District and Bellevue Public School District and consisted of children in first grade (n = 18) and second grade (n = 13). Participants were not excluded if they were receiving special education services, provided they also participated within the regular education classroom. Inclusion criteria were (a) the child had to be identified by a teacher or therapist as having a handwriting performance skill deficit in the classroom; (b) spoken language had to be the child's primary means of communication; and (c) English had to be the primary language spoken in the home. Further, each child had to be able to sustain attention for a 20-min seated activity as verified by his or her teacher. The two school districts were chosen in order to include variety in socioeconomic and ethnic background. Before initiation of the study, human subjects approval was obtained from the University of Washington Human Subject Committee and from the participating school districts. Participants belonged to the following ethnic groups: 15 Caucasian, 7 African-American, 3 Hispanic, 3 Asian, 2 Native American, and 1 East Indian.

Instrument

The ETCH-M (Amundson, 1995) was designed to evaluate a student's handwriting legibility and rate of writing. It requires the student to complete a variety of writing tasks similar to those expected in a classroom. It evaluates six areas of handwriting: (a) alphabet production in uppercase and lowercase from memory, (b) numeral writing from memory, (c) near-point copying, (d) far-point copying, (e)
Table 1  
Task Descriptions and Potential Raw Score Ranges on the ETCH-M

<table>
<thead>
<tr>
<th>Task Description</th>
<th>Potential Raw Score Range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Letters</td>
</tr>
<tr>
<td>ia. Alphabet lowercase (a to z) from memory</td>
<td>0-26</td>
</tr>
<tr>
<td>ib. Alphabet uppercase (A to Z) from memory</td>
<td>0-26</td>
</tr>
<tr>
<td>ii. Numeral writing (1 to 12) from memory</td>
<td>NA</td>
</tr>
<tr>
<td>iii. Near-point copy (5 words, 18 letters) from task sheet</td>
<td>0-18</td>
</tr>
<tr>
<td>iv. Far-point copy (5 words, 18 letters) from wall chart</td>
<td>0-18</td>
</tr>
<tr>
<td>v. Dictation (15 letters and numerals total) by verbal direction</td>
<td>0-10</td>
</tr>
<tr>
<td>vi. Sentence composition (5 or more words) creative writing</td>
<td>Varies</td>
</tr>
</tbody>
</table>

Note: All raw scores are converted to percentage equivalents for data analysis (0-100%). All tasks are timed and recorded in letters per minute, except dictation, which is not timed. ETCH-M = Evaluation Tool for Children’s Handwriting-Manuscript; NA = not applicable.

dictation, and (f) sentence composition. This criterion-referenced tool has standard administration and scoring procedures, is designed to be administered individually, and requires 20 min to 30 min to complete. Scoring focuses on legibility, rate of writing, components of legibility, and biomechanical aspects of handwriting. For legibility scoring, if a letter, number, or word is individually determined to be readable, it is counted as a correct item on the score sheet. The number of illegible items is subtracted from the total number of items. The difference is then converted to a percentage score (see Table 1 for task descriptions and potential ranges of raw scores). Qualitative information is recorded on the basis of the legibility components, including letter formation, letter size, horizontal alignment, and spacing. The score sheet also allows for data collection related to the student’s mechanical aspects of pencil grasp pattern, pencil position, pencil pressure, and in-hand manipulation of the pencil along with classroom observations. For purposes of this study, only the specific legibility percentages were evaluated for test-retest reliability.

A chapter in the ETCH examiner’s manual assists the examiner in learning the complex scoring criteria before scoring actual test protocols (Amundson, 1995). A self-study tutorial and scoring competency quizzes provide training in scoring. An examiner is expected to achieve a scoring competency of 90% on the trial tests in the manual before attempting to score children’s performances in the practice setting.

Procedure

Each student was administered the ETCH-M by the primary investigator, who followed the examiner’s manual guidelines. Before the study, the primary investigator administered the ETCH-M to five early elementary-school-age children to gain experience in administering the test, but these children were not included in the study. To further control for potential error in test administration, a test procedural agreement checklist was developed on the basis of the standardized procedures in the ETCH examiner’s manual. During every 10th or 12th test, a second person sat in an unobtrusive location during administration and recorded procedural correctness with the procedural agreement checklist. There was a total of five procedural checks with procedural agreement, ranging from 91% to 100%.

The ETCH-M was administered to the 31 participants on two separate occasions, using a 7-day time interval. Anastasi (1988) reported that short intervals between test and retest are appropriate with young children for whom discernible developmental changes can be expected within a limited time frame. The study began during the second half of the school year so that the students in first grade would be exposed to the teaching of the entire alphabet in the classroom. The test was administered individually in a quiet area of each participant’s school between 8:30 a.m. and 12:00 p.m. The participant was seated at the same desk for test and retest. The participant used a sharpened standard number 2 pencil with eraser during the test and received a colorful pencil after each testing. Each writing sample was coded with a number. Because testing occurred over a span of 2 months, a low code number could be a test or a retest score sheet.

The primary investigator acted as the sole rater of the 62 score sheets and was blind both to the participants’ identities and to whether a particular score sheet reflected a test or retest administration. She scored the test score sheets in randomly selected groups of 10 after all 62 tests had been administered. The fact that a code number could be high or low for test and retest and the addition of time between test administration and test rating combined to reduce the risk of the primary investigator’s ability to identify the participant.

Before rating the score sheets, and periodically throughout the rating process (after every 10th score sheet), the primary investigator checked scoring competency according to the examiner’s manual instructions. In all cases, her scoring competency was at or above the 90% standard described in the manual. This periodic retesting of competency was an attempt to control rater drift (Goldstein & Tupper, 1987).

Data Analysis

Descriptive statistics were examined for test and retest for individual tasks and total scores. Raw scores were converted to percentages as instructed in the examiner’s manual. Next, for each individual writing task and the total legibility scores, the use of parametric versus nonparametric
statistics was determined after examining scatterplots, histograms, measures of central tendency and score variability, and kurtosis and skewness scores. When data were approximately normally distributed and other assumptions for the use of parametric statistics were met, intraclass correlation coefficients (ICCs) were computed. The ICC was chosen because it is the preferred statistic for estimating the reliability of repeated measures in addition to being sensitive to systematic measurement error when data are approximately normally distributed (Baumgartner, 1989). In situations where data did not meet the assumptions for use of parametric statistics, Spearman rank order correlations were used. It is possible with Spearman rank order correlations to have a high correlation between test and retest when test scores are either consistently higher or consistently lower than retest scores. To check for such systematic differences between test and retest scores, Wilcoxon signed rank tests were used. For both the Spearman rank order correlations and ICC, a test–retest reliability coefficient at or above .80 was chosen as the desired level.

Finally, test–retest agreement for individual and total legibility scores was addressed through magnitudes of difference between test and retest scores. This was done by first subtracting retest percentage scores on the ETCH-M from test percentage scores. Then, the number of scores for each legibility task reflecting no change and specific magnitudes of either positive or negative change were reported. A positive change might reflect a practice or learning effect, whereas a negative change might suggest that students became bored with the test on second administration.

Results
Descriptive statistics for letter legibility of individual tasks are reported in Table 2, and letter, numeral, and word legibility of total scores are reported in Table 3. Reliability coefficients for the individual tasks ranged from .20 for the near-point copy task to .76 for the alphabet uppercase task. For the total scores, the reliability coefficients ranged from .63 for the total numeral legibility task to .77 for the total letter legibility with composition task. For those tasks where it was necessary to use Spearman rank order correlations because the data were not approximately normally distributed, analyses were done to check for systematic error. First, examination of differences between means of test and retest outcomes indicated that there was no improvement of greater than .06 on the individual writing tasks and .02 on the total legibility scores. For three individual writing tasks, the mean score stayed the same (near-point copy) or decreased slightly on the retest (numeral writing, far-point copy). In addition, two total score means decreased (total letter legibility excluding the composition task, total numeral legibility). The maximum change between test and retest means on the individual tasks was .06 for dictation. Second, according to Wilcoxon signed rank tests, no significant differences were found between test and retest scores.

Magnitudes of difference for the letter legibility of individual tasks and letter, numeral, and word legibility of total tasks with and without the composition task are reported in Table 4. An example of how to read and interpret Table 4 is as follows: For total letters, the findings indicated that 1 participant's scores did not change between test and retest; 14 participants' scores increased by 1% to 10%; 10 participants' scores decreased by 1% to 10%; and 3 participants' scores either increased or decreased by 11% to 20%.

Discussion
Overall, test–retest reliability of handwriting legibility as measured by the ETCH-M, using a 7-day interval, was lower than desirable for this group of first-grade and second-grade students with identified handwriting deficits. Reliability coefficients were .77 for total letter legibility, .71 for total word legibility, and .63 for total numeral legibility. Individual task reliability coefficients were generally lower and ranged from .20 (near-point copy) to .76 (alphabet uppercase). With the exception of alphabet uppercase, individual task scores appeared less stable over time than did the total numeral and total word legibility scores. Therefore, these individual task scores are not recommended for use in determining eligibility for services or in documenting change over time.

Results were reported with the ICC for those data that were approximately normally distributed. The ICC is sensitive to systematic error (e.g., systematic improvements in scores due to maturation or practice). Because this is a more robust statistic than the Pearson product–moment correlation, the reliability outcomes in this study may appear lower than those in similar studies where other statistics were used.

The magnitudes of difference (see Table 4) illustrate two strengths. First, results revealed no substantial practice effect overall. Second, no test–retest scores changed more than 20% for total letter legibility scores. This suggests that if a child's total letter legibility score changes more than 20% between test and retest, then it likely is attributable to clinical change and not measurement error. In contrast, the individual task of sentence composition showed the highest score variability, with 32% of the participants changing more than 20% between test and retest. This large variability between test and retest on the composition task may be due to the complicated and spontaneous nature of being asked to generate a five-plus word sentence. The scoring of legibility for this task is
Table 2
Descriptive Statistics and Test-Retest Reliability Coefficients for ETCH-M Individual Writing Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>$M$</th>
<th>Median</th>
<th>$SD$</th>
<th>Low Score/High Score</th>
<th>Reliability Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. Alphabet lowercase</td>
<td>.73</td>
<td>.81</td>
<td>.20</td>
<td>.12/.100</td>
<td>.64&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Test</td>
<td>.74</td>
<td>.77</td>
<td>.18</td>
<td>.15/.03</td>
<td></td>
</tr>
<tr>
<td>Retest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1b. Alphabet uppercase</td>
<td>.69</td>
<td>.69</td>
<td>.18</td>
<td>.23/.100</td>
<td>.76&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Test</td>
<td>.72</td>
<td>.77</td>
<td>.19</td>
<td>.12/.100</td>
<td></td>
</tr>
<tr>
<td>Retest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II. Numerical writing</td>
<td>.85</td>
<td>.92</td>
<td>.14</td>
<td>.50/.100</td>
<td>.63&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Test</td>
<td>.83</td>
<td>.92</td>
<td>.18</td>
<td>.33/.100</td>
<td></td>
</tr>
<tr>
<td>Retest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>III. Near-point copy</td>
<td>.88</td>
<td>.89</td>
<td>.09</td>
<td>.67/.100</td>
<td>.20&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Test</td>
<td>.88</td>
<td>.89</td>
<td>.11</td>
<td>.56/.100</td>
<td></td>
</tr>
<tr>
<td>Retest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV. Far-point copy</td>
<td>.88</td>
<td>.89</td>
<td>.11</td>
<td>.61/.100</td>
<td>.63&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Test</td>
<td>.86</td>
<td>.89</td>
<td>.13</td>
<td>.56/.100</td>
<td></td>
</tr>
<tr>
<td>Retest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V. Dictation</td>
<td>.70</td>
<td>.73</td>
<td>.23</td>
<td>.13/.100</td>
<td>.68&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Test</td>
<td>.76</td>
<td>.80</td>
<td>.22</td>
<td>.00/.100</td>
<td></td>
</tr>
<tr>
<td>Retest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI. Sentence composition</td>
<td>.74</td>
<td>.78</td>
<td>.21</td>
<td>.06/.100</td>
<td>.33&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Test</td>
<td>.79</td>
<td>.80</td>
<td>.23</td>
<td>.00/.100</td>
<td></td>
</tr>
<tr>
<td>Retest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. All descriptive statistics are converted from raw scores and reported in percentage format. The above tasks refer to individual letter or number legibility, not word legibility. ETCH-M = Evaluation Tool for Children's Handwriting-Manuscript.
<sup>a</sup>Spearman rank order correlation coefficient. <sup>b</sup>Intra class correlation coefficient.

also more complex because of the high degree of variation of sentences and words chosen by each participant.

Test–retest reliability for the ETCH-M was lower than desired for test development but within the range of other assessment tools measuring children's handwriting performance. Using a handwriting assessment tool they developed, Ziviani and Elkins (1984) reported a range from .48 to .84 of test–retest reliability when examining the handwriting of 575 Australian children between 7 and 14 years of age in regular education classrooms. On the Minnesota Handwriting Test-Research Version, Reisman (1993) found a range of .58 to .94 for test–retest reliability on a convenience sample of students in Maine, Wisconsin, and Minnesota. Neither study examined children with identified handwriting problems, and each used the Pearson product–moment correlation.

Several factors may explain the low test–retest reliability coefficients of this study. First, ceiling effects may have had an impact on some individual task scores and total scores. Twenty percent or more of the participants scored 100% on either the test or retest for three individual task scores (near-point copy, far-point copy, numeral writing) and for one total task score (total numeral legibility).

Second, the scoring of handwriting legibility is inherently subjective. Whereas some other handwriting tools

Table 3
Descriptive Statistics and Test–Retest Reliability Coefficients for Total Scores on the ETCH-M

<table>
<thead>
<tr>
<th>Letter and Numeral Total</th>
<th>$n$</th>
<th>$M$</th>
<th>Median</th>
<th>$SD$</th>
<th>Low Score/High Score</th>
<th>Reliability Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total letter legibility, excluding Task VI&lt;sup&gt;c&lt;/sup&gt;</td>
<td>31</td>
<td>.69</td>
<td>.71</td>
<td>.75</td>
<td>.24</td>
<td>.11/.100</td>
</tr>
<tr>
<td>Test</td>
<td>31</td>
<td>.78</td>
<td>.81</td>
<td>.14</td>
<td>.25</td>
<td>.08/.100</td>
</tr>
<tr>
<td>Retest</td>
<td>28</td>
<td>.78</td>
<td>.80</td>
<td>.13</td>
<td>.35/0.97</td>
<td>.71&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Total numeral legibility</td>
<td>31</td>
<td>.79</td>
<td>.81</td>
<td>.10</td>
<td>.36/0.97</td>
<td>.77&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Test</td>
<td>31</td>
<td>.83</td>
<td>.79</td>
<td>.11</td>
<td>.53/0.96</td>
<td>.77&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Retest</td>
<td>28</td>
<td>.81</td>
<td>.88</td>
<td>.15</td>
<td>.50/0.99</td>
<td>.77&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Total word legibility</td>
<td>28</td>
<td>.67</td>
<td>.73</td>
<td>.20</td>
<td>.47/1.00</td>
<td>.63&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Test</td>
<td>28</td>
<td>.69</td>
<td>.76</td>
<td>.23</td>
<td>.35/1.00</td>
<td>.63&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Retest</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. All descriptive statistics are converted from raw scores and reported in percentage format. ETCH-M = Evaluation Tool for Children's Handwriting-Manuscript.
<sup>a</sup>Spearman rank order correlation coefficient. <sup>b</sup>Intra class correlation coefficient. <sup>c</sup>Task VI is the sentence composition task that three children refused to perform, thus lowering the sample size to 28.
have more detailed analytical rating systems requiring transparent overlays, rulers, and complex scoring criteria for each individual letter’s legibility (Reisman, 1993; Tseng & Cermak, 1991; Ziviani & Elkins, 1984), the ETCH focuses on legibility with a more global measurement method. The ETCH targets legibility in terms of readability of the student’s handwriting, supported by specific scoring criteria of letters, numerals, and words. To determine letter legibility, the therapist must compare the child’s handwriting to illegible and legible models found in the ETCH examiner’s manual. The handwriting models do not cover the inexhaustible number of possible letter formations produced by children; thus, the examiner is required to score the letter as correct if there is any question regarding its being legible. Because of the nature of this scoring method, subjectivity increases, and a greater possibility for inconsistencies in scoring exist. In this study, the primary investigator scored the entire sample in an attempt to control for scoring variability.

Third, because of the complex nature of handwriting skills, the quality of a child’s handwriting performance tends to vary over time. The development of handwriting ability involves the refinement, synthesis, and integration of sensorimotor, cognitive, and language abilities (Ziviani, 1995) and is one of the more complex tasks performed by young children as they enter school. This complexity can result in substantial variations in handwriting performance from day to day for the same child (Herrick, 1960). In addition, children referred to occupational therapy are usually suspect of an unstable or immature neurological system, which can add even more discrepancy between their handwriting performances over time. In the current study, an effort to minimize the influence of environmental factors was made by controlling the time of day, day of week, and test location.

Fourth, in the retest session, some participants appeared to begin the test before listening to the complete directions, and a few resisted performing the retest. These situations could have interfered with achieving the participants’ best efforts on the retest. Although no predictable rise or fall occurred, retesting could have affected some participants’ performance.

Fifth, the sample of children identified to have handwriting deficits likely included those with varying degrees of attention deficits, sensory processing problems, neuromuscular impairments, auditory processing problems, language difficulties, and learning deficits. Children with these underlying problems are likely to have difficulty in following instructions, attending to task, and completing tasks; all are necessary for testing performance. This study attempted to control for attention deficits by specifying within the inclusion criteria that the child must be able to sustain attention for a 20-min seated activity. Even so, there may have been unidentified subgroups in the sample who performed better or worse on the retest because of attention difficulties. Although inclusion of these children may have compromised the results, their inclusion was appropriate because the ETCH was designed for this target population.

Finally, young children exhibit a high level of variability in performance, particularly in early elementary school years when handwriting is being taught and mastered. Ziviani and Elkins (1984) found that younger children (grade 3) demonstrated lower test–retest correlations than older children (grade 6). They concluded that older children generally show more consistency in handwriting legibility over time than do younger children. Because of the high referral rate of early elementary school children to occupational therapy, the variability of their performance, and the target population of the ETCH-M, first-grade and second-grade students were selected for the current study.

During testing and retesting, a few participants appeared to have difficulty with completing the composition task. Several were unable to produce words in a sentence format, and three refused to attempt the task.
claiming that it was "too hard." Although sentence composition is commonly required in the classroom, these participants' inability to produce words could have affected test–retest reliability outcomes. Total scores were compared with and without the composition task and are reported in Table 4. The correlation coefficients did not vary sufficiently in the combined first-grade and second-grade population to support the hypothesis that the sentence composition task lowered the total scores.

In addition to the legibility percentage scores examined here, the ETCH also examines writing speed, the impact of legibility components (i.e., spacing, size, horizontal alignment, letter formation) on readability, pencil manipulation, hand preference, grasp pattern, pencil management, and classroom observation. Recommendations for intervention evolve from a combination of these factors. Although scores from a standardized test are useful, the occupational therapist gathers information from several sources before recommending intervention.

Clinical Implications

When evaluating a student's handwriting, it is important to consider ETCH legibility scores as only one aspect of a comprehensive evaluation. Direct classroom observation of the student involved in a writing assignment, parent and teacher discussions, classroom samples, and an educational cumulative file review are imperative to view the student's occupation of writing in context. The use of the ETCH and other data-gathering techniques may allow the team to gain a comprehensive picture of a student's handwriting abilities as well as formulate a blueprint for intervention.

Study Strengths and Limitations

This study had several strengths. First, the sample represented the target population of the ETCH: children with handwriting problems. Second, the sample consisted of a heterogeneous ethnic mix. Third, time of day, test location, and a 7-day interval were consistent for test and retest for each participant. Fourth, one person administered and scored the tests to maintain consistency. Fifth, steps also were taken to minimize score variance due to rater error. When the primary investigator scored each test, she was blind to the participant's identity and whether a particular score sheet was from a test or retest administration. Additionally, she rechecked scoring competency periodically throughout the study.

The limitations to the study were the small sample size and small number of school districts used. However, the ethnic mix and proportionally high number of classrooms from which children were chosen help to increase the generalizability of the results.

Directions for Future Research

Replication of this study with a larger sample would improve generalizability. A comparison study is needed of the test–retest reliabilities for first-grade students with those for second-grade students on the ETCH-M because of the apparent variability in early elementary school years. A reliability study of the cursive segment of the ETCH also is needed. Possibly another test–retest reliability study could address the legibility components (i.e., spacing, horizontal alignment, letter formation, size), the rate of writing, and the biomechanical aspects of handwriting as assessed by this tool.

Conclusion

Findings from this study suggest that caution should be used when reporting legibility percentages of the ETCH-M, specifically with first-grade and second-grade students with identified handwriting deficits. As with other handwriting assessment tools, test–retest reliability is lower than desired for pediatric standardized tests. The complexity of handwriting for early elementary school children, the subjectivity of readability, and the high variability of performance among children with immature or unstable neurological systems may contribute to the instability of handwriting performance over time. Percentage scores for total letter legibility, total word legibility, and uppercase letter legibility were more stable than total numeral legibility scores and legibility for the other individual tasks. Therapists are encouraged to use total letter and total word percentages as the most reliable scores over time for the combined first-grade and second-grade population. To reflect "best pediatric practice," standardized testing should be combined with contextual data-gathering techniques.

Acknowledgments

We thank the students, parents, teachers, and occupational therapists of the Bellevue and Seattle Public School Districts who assisted and participated in the study. This study was supported, in part, by Grant No. 84.029F, "Preparation of Related Service Personnel: Preservice Training of Occupational Therapy to Provide Services to Children with Emotional and Behavioral Disorders," from the U.S. Department of Education, Office of Special Education and Rehabilitative Services.

This study was conducted in partial fulfillment of the first author's requirements for a master of science degree, Department of Rehabilitation Medicine, University of Washington, Seattle, Washington.

References


The Essence of play: A Child’s Occupation

Edited by Barbara E. Chandler, MOT, OTR

This interesting explanation of the nature of play includes theoretical and practical applications for occupational therapy practitioners. Individual chapters by experts help readers understand play as a universal occupation and how it can be used as a therapeutic medium for infants and children. Presents play from a neurological viewpoint as well as an anthropological perspective written by well-known anthropologists Wendy Wood, PhD, OTR/L, and Reba Anderson, PhD, OTR, FAOTA. 220 pages, indexed, 1997.

Table of Contents

- Play as an Occupation and Indicator of Health
- Insights From the Play of Nonhuman Primates
- The Anthropological Study of Play
- Infant Play
- What Happens When We Play?
  A Neurodevelopmental Explanation
- Play Things: Toy Use, Accessibility, and Adaptation
- Playmates: Social Interaction in Early and Middle Childhood
- Play Environments: An Occupational Therapy Perspective
- Play as Treatment and Treatment Through Play

Order #1184 $35 AOTA member $45 nonmember

To order, call 1-800-SAY-AOTA (AOTA members), 301-652-2682 (nonmembers), or 1-800-377-8555 (TDD users). Shipping and handling additional.

The American Journal of Occupational Therapy