Interrater Reliability of a New Handwriting Assessment Battery for Adults

Kathrine Faddy, Annie McCluskey, Natasha A. Lannin

Objective. The objective of this study was to develop, pilot, and evaluate the interrater reliability of a new Handwriting Assessment Battery for adults.

Design. Test development included item selection and interrater reliability involving two raters.

Method. The test assessed pen control and manipulation, writing speed, and writing legibility. Ten people with brain injury completed the test with two occupational therapists independently rating 10 writing samples. Results were analyzed for reliability using kappa and intraclass correlation coefficients (ICC).

Results. Pen control and manipulation subtests showed high to perfect agreement (line drawing subtest, $\kappa$ = 1.0; dot subtest, $\kappa$ = 0.80). The speed subtest showed perfect agreement (ICC = 1.0). Writing legibility showed high agreement for all five subtests (ICC = 0.71–0.83), although a ceiling effect was evident for two subtests.

Conclusion. Although the test showed excellent interrater reliability, further reliability and validity testing are needed before the test is used clinically.

Handwriting is an important occupational task. Adults need to be able to write to sign their names, record messages, and complete shopping lists. Adults with a stroke or brain injury often require handwriting retraining. Yet no comprehensive handwriting assessment exists for use with adults to guide therapists during rehabilitation.

Typical writing problems after acquired brain injury are reported to include difficulty picking up and manipulating a pen and generating adequate muscle force and pressure to write (Carr & Shepherd, 2003; Miles Breslin & Exner, 1998). Pen control is important for precise written output. A study of 48 first-grade students (Cornhill & Case-Smith, 1996) found that in-hand manipulation and pen control were associated with improved letter formation and may, therefore, be associated with improved handwriting legibility. Although a dynamic tripod grip has traditionally been thought to lead to faster, more legible writing (Tseng & Cermak, 1993), this assumption has not been confirmed through research. Speed and legibility were not affected even by the most atypical grip patterns in a study involving 282 children up to age 14 (Ziviani & Elkins, 1986).

Speed is important for functional handwriting because text needs to be written in a reasonable time (Rosenblum, Weiss, & Parush, 2003). Slow writing speed is one reason for referral to occupational therapy in the adult population. Poor legibility, however, is probably the most common reason for referral in this population, at least anecdotally and in the absence of published data. We have collectively seen many people in rehabilitation with stroke and brain injury who complain of untidy, illegible writing. If legibility is poor, research involving children has found that
letters, numbers, and words may be unrecognizable to anyone but the writer (Diekema, Deitz, & Amundson, 1998). Perceptual–motor skills, particularly hand–eye coordination, are often the focus of pediatric handwriting training programs (Graham & Weintraub, 1996) but anecdotally are less commonly a problem for adults who learn to write before their stroke or brain injury.

Existing handwriting assessments have been validated for use with children (Amundson, 2002; Feder & Majnemer, 2003; Reisman, 1993; Wallen, Bonney, & Lennox, 1996). One test, the Evaluation Tool of Children’s Handwriting (ETCH; Amundson, 1995), appeared to be suitable for use and adaptation with adults. Although designed to measure writing legibility in young children (Feder & Majnemer, 2003), the subtests contain characters that both children and adults write. For example, participants are required to write letters of the alphabet (printed and cursive), numbers, and a self-generated sentence. Thus, several subtests of the ETCH could be adapted for use with adults.

In addition to adapting a pediatric test, we reviewed relevant subtests of existing upper-limb assessments. One such test was the Motor Assessment Scale (MAS; Carr, Shepherd, Nordholm, & Lynne, 1985) for use with adults after a stroke. The MAS includes two tests of pen control, requiring a person to draw 10 lines and 10 dots in a specified time. Another assessment, the Jebsen-Taylor Test of Hand Function (Jebsen, Taylor, Trieschmann, Trotter, & Howard, 1969) includes seven subtests, one of which evaluates writing speed. Participants are asked to copy a sentence of 24 letters as quickly as possible. The written output is timed. Using subtests from one pediatric and these two adult upper-limb assessments, McCluskey and Lannin (2003) developed the Handwriting Assessment Battery (HAB). The aims of this study were to develop, pilot, and explore the interrater reliability of the HAB.

Method

Participants

Ten adults with brain injury and two occupational therapy raters participated. A convenience sample was recruited of 10 adults with traumatic brain injury (6 men, 4 women) who self-identified as having difficulty with writing. Eligibility criteria included being age 18 years or older, living in the Sydney metropolitan area, having a diagnosis of traumatic brain injury, being at least 1 year after injury, having difficulty holding or using a pen, and being able to attempt to write unprompted. The latter two criteria excluded people who could not hold a pen and those with apraxia or planning problems. Ethical approval was obtained from a relevant university committee, and written consent was provided by each participant before data collection. The two raters each had more than 10 years of experience. Although both had used standardized tests, neither rater had used any of the HAB subtests before.

Description of the HAB

The HAB contains three sections with items selected and adapted from eight subtests of the MAS (Carr et al., 1985), the Jebsen (Jebsen et al., 1969), and the ETCH (Amundson, 1995). Each of the HAB subtests provides a profile of performance in the areas of pen control and manipulation, writing speed, and writing legibility. There is no total summed score. The HAB takes approximately 20 min to administer and 15 min to score.

Section 1 of the HAB (Pen Control and Manipulation) contains two subtests from the MAS (Carr et al., 1985). These subtests require a person to (1) draw at least 10 horizontal lines with a pencil in 20 s on an A4 page with two premarked vertical lines, with the person starting and stopping at the vertical lines, and (2) make 10 or more rapid consecutive dots with a pencil in 5 s. Participants are scored on the best of three attempts. Each subtest is scored as either “Achieved” (score of 1) or “Not achieved” (score of 0).

Section 2 of the HAB (Writing Speed) contains one subtest and is derived from the Jebsen-Taylor Test of Hand Function (Jebsen et al., 1969). This subtest involves the timed copying of a sentence of third-grade reading difficulty using the dominant hand only. One prewritten sentence is selected randomly from a set of three (Jebsen et al., 1969), each containing 24 letters. The three sentences are as follows: “The old man seemed to be tired,” “Fish take air out of the water,” and “John saw the red truck coming.” Participants copy the sentence as quickly as possible using printed or cursive writing. The time, in seconds, is recorded and compared with published norms for adults ages 16 to 25 (Agnew & Maas, 1982) and 20 to 94 (Jebsen et al., 1969).

Section 3 of the HAB (Writing Legibility) contains five legibility subtests derived from the ETCH (Amundson, 2002): (1) lowercase alphabet (a to z); (2) uppercase legibility (A to Z); (3) numeral legibility (numbers 1–12); (4) sentence composition–words; and (5) sentence composition–letters. Each subtest provides a percentage legibility score, which is determined by comparing letters and words to acceptable or unacceptable examples in the training manual. The examples help raters to identify legible or illegible letters. The maximum legibility score is 100% (Feder & Majnemer, 2003).
A HAB training manual (Faddy, 2004) describes the standardized administration of subtests, equipment required, and how to score subtests. Examples of acceptable or unacceptable responses for the dot, line drawing, and legibility subtests are provided.

**Procedure for HAB Data Collection and Rater Training**

Data were collected from the 10 people with brain injury on a single occasion by Kathrine Faddy at the person’s home or local community center. With consent, video footage was obtained during each test. The 10 deidentified writing tests and videos were copied for each rater. Raters attended a 3-hr workshop where they practiced scoring a test video and scored sample tests using the training manual. Raters then scored the 10 writing tests independently over 2 weeks. Scoring took raters up to 15 min per test.

**Data Analysis**

Kappa and intraclass correlation coefficients (ICCs) were used. For dichotomous data from Section 1 (Achieved or Not Achieved), the kappa coefficient was used to represent raters’ agreement. For continuous data (time and percentage legibility) produced from Sections 2 and 3, ICC_{2,1} was used to represent raters’ agreement. In exploratory studies such as this, a cut-off coefficient value of 0.70 was considered acceptable (Polit, Beck, & Hungler, 2001).

**Results**

The mean time after injury of people with brain injury in this study was 4.6 years (range = 1 to 18); mean age at entry to the study was 34.9 years (range = 21 to 49).

For Section 1, there was excellent (κ = 1.0, p = .010) and very good (κ = 0.80, p = .002) agreement between raters for the horizontal line and dot subtest scores, respectively. For Section 2, there was excellent agreement for writing speed scores (ICC = 1.00), with narrow confidence intervals (95% CI = 0.99–1.00, p = .0001). Writing speed scores ranged from 27 to 174 s.

For Section 3, there was a high correlation and agreement between raters, with ICC values from 0.71 (Sentence Composition–Letters) to 0.83 (Numeral Legibility). The ICCs, CIs, p values, and interpretation of ICC values for Section 3 are presented in Table 1. Percentage legibility scores ranged from 0% (illegible) to 100% (fully legible). Of the 100 legibility subtests scored by raters (5 subtests × 10 samples scored per rater = 50 samples per rater), only 6/100 samples were scored as having 0% legibility across three subtests: Lower Case Legibility (n = 1), Numeral Legibility (n = 1), and Sentence Composition–Words (n = 4). Of the 100 legibility subtests scored by raters, 21/100 were scored as having 100% legibility across four subtests: Upper Case Legibility (n = 3), Numeral Legibility (n = 7), Sentence Composition–Words (n = 10), Sentence Composition–Letters (n = 1).

**Discussion**

The high to very high agreement between raters in this study suggests that adaptation and use of the MAS, Jebsen, and ETCH subtests did not negatively affect scoring. Comparable results were found to those reported by Carr and colleagues (1985) and more recently by Lannin (2004) for the MAS subtests involving pen control. The speed subtest taken from the Jebsen, and writing legibility subtests derived from the ETCH, produced higher correlation values and rater agreement than reported by the original authors (Amundson, 1995; Feder & Majnemer, 2003; Jebsen et al., 1969; Sudsawad, Trombly, Henderson, & Tickle-Degnen, 2002).

Coefficients for legibility subtests of the HAB were higher (ICC = 0.71–0.83) than those reported by Amundson (1995; ICC = 0.42–0.88). These differences may have occurred because our raters received face-to-face training at a workshop. They practiced scoring sample tests and discussed what

---

**Table 1. Interrater Reliability for Section 3 (Writing Legibility) of the Handwriting Assessment Battery**

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Measure of Agreement</th>
<th>95% CI of ICC</th>
<th>Significance (p value)</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Case Legibility</td>
<td>0.78</td>
<td>0.05–0.95</td>
<td>.0001</td>
<td>High correlation</td>
</tr>
<tr>
<td>Upper Case Legibility</td>
<td>0.81</td>
<td>0.61–0.96</td>
<td>.0001</td>
<td>High correlation</td>
</tr>
<tr>
<td>Numeral Legibility</td>
<td>0.83</td>
<td>0.26–0.96</td>
<td>.0001</td>
<td>High correlation</td>
</tr>
<tr>
<td>Sentence Composition–Words</td>
<td>0.79</td>
<td>0.38–0.94</td>
<td>.001</td>
<td>High correlation</td>
</tr>
<tr>
<td>Sentence Composition–Letters</td>
<td>0.71</td>
<td>0.10–0.92</td>
<td>.002</td>
<td>High correlation</td>
</tr>
</tbody>
</table>

*Note. CI = confidence interval.*

aMeasure of agreement = intraclass correlation coefficients (ICC)

b*p < .05.
were considered legible (acceptable) and illegible (unacceptable) letters. However, raters in the Amundson (1995) study only read the *ETCH Examiners Manual*. Training raters together at a workshop or using an interactive online learning package may be necessary to achieve good agreement between raters for legibility (Moon & Hughes, 2002). Another reason for higher levels of agreement when scoring legibility may be that the HAB does not require raters to differentiate between manuscript and cursive text. Unlike the ETCH, if a writer mixes uppercase and lowercase text (manuscript and cursive writing) together in a sentence, they do not lose marks. We noted during pilot testing that adults commonly mix the two types of text and considered this acceptable. Further observations with a larger sample of adult writing are needed to confirm whether most adults write in this style.

In this reliability study, a negligible floor effect was found: Only 6% (*n* = 6) of writing samples scored 0% for the legibility subtests. However, a ceiling effect was evident, because 21% (*n* = 21) of writing samples scored 100% for legibility, the maximum possible score. A ceiling effect was observed predominately in two subtests: Numeral Legibility (*n* = 7) and Sentence Composition–Words (*n* = 10). These subtests may not be sufficiently challenging to discriminate between legible and illegible handwriting in this population. The main implication of a ceiling effect is that the HAB may fail to detect clinical improvements in people referred for handwriting training if they score 100% at baseline.

Further studies are needed to establish content and construct validity and test–retest reliability. A larger sample would also allow Rasch modeling to examine the item performance, particularly in light of the potential ceiling effects of the Numeral Legibility and Sentence Composition–Words subtests. Normative data are also needed to provide information about the handwriting performance of adults across ages and occupations. Anecdotally, adults write less and correspond more using mobile phones, text messaging, and e-mail than in previous decades; this change in communication methods has the potential to reduce handwriting speed and legibility. A random selection of participants, as well as a larger sample, would also be preferable in future studies. The nature and size of our sample were limitations to the generalizability of results. Although the sample was deemed to be typical of adults with traumatic brain injury, writing samples from other disability groups may be easier or more difficult for raters to score. Future studies should also recruit people with other conditions such as stroke, multiple sclerosis, Parkinson’s disease, and occupational overuse syndrome.

In conclusion, the HAB was designed to measure handwriting retraining. Individual subtests produced high to perfect interrater agreement. Because no other comprehensive and standardized adult handwriting assessment exists, further validity and reliability testing of the HAB is warranted. Normative data also need to be collected, against which the performance of rehabilitation populations can be compared. Currently, the HAB is a valuable tool for therapists to measure outcomes of handwriting retraining with patients after acquired brain injury until a more comprehensive assessment is developed or further study is completed on the HAB.

### References


New South Wales, School of Exercise and Health Sciences, University of Western Sydney.