A Comparison of Computerized and Paper-Based Language Tests With Adults With Aphasia

Caroline Newton, a Kadia Acres, b and Carolyn Bruce a

Purpose: This study investigated whether computers are a useful tool in the assessment of people with aphasia (PWA). Computerized and traditionally administered versions of tasks were compared to determine whether (a) the scores were equivalent, (b) the administration was comparable, (c) variables such as age affected performance, and (d) the participants’ perceptions of the computerized and traditionally administered versions of the tasks were similar.

Method: Fifteen PWA were assessed on 2 language tasks—sentence-picture matching and grammaticality judgment—in 3 conditions: computer only, computer with the clinician present, and traditional. The participants also completed questionnaires rating aspects of each condition.

Results: Scores from the traditionally administered tasks were highly correlated with those from the computerized tasks, but scores from the computerized tasks were significantly lower. There was no significant difference in the time taken between the conditions. Whereas some individuals felt comfortable with the computer, overall, participants preferred the traditional assessment method or when another person was in the room. No factors were identified that predicted participants’ relative performance in the computer condition.

Conclusion: The results suggest that PWA can be assessed using computerized versions of tasks, but that caution should be exercised when comparing scores to those collected using traditional methods, including norms. The variation in participants’ opinions regarding computerized tasks suggests that this method might be more suitable for some participants than others.

Key Words: aphasia, assessment, computers, language

The use of technology in the assessment and management of speech and language difficulties has been a major theme in speech-language pathology for many years, and with frequent new developments and innovations, it should continue to increase for some time to come. One area that has seen an increased use of computers is aphasia treatment (e.g., Mortley, Wade, Enderby, & Hughes, 2004; Palmer, Enderby, & Hawley, 2007; Ramsberger & Marie, 2007; Wertz & Katz, 2004). In a survey of U.S. speech-language pathologists (SLPs), Davis and Copeland (2006) found that >50% of respondents used computers in aphasia treatment, including treatments targeting reading, writing, and word-finding difficulties. Using computers to assess aphasia is less common, but with increasing pressures on SLP services in terms of both staff numbers and staff time, computerized assessment may have great potential to relieve some of those pressures and bring both financial and administrative advantages. For example, the use of video-conferencing for the remote assessment of clients who live in rural areas, where SLP services may be particularly under-resourced, has been shown to provide reliable diagnoses of speech and language disorders (Mashima & Doarn, 2008; Theodorus, Hill, Russell, Ward, & Wootton, 2008).

There are also many potential advantages for assessment where the client inputs his or her own responses into the computer either in the clinic or online, including the fact that this could be carried out without an SLP present. Other advantages of computerized assessment include improved reliability and objectivity, decreased scoring errors, and increased opportunity to process the data online (e.g., summarizing scores, automatically selecting items; Schulenberg & Yutrzenka, 2004). Computerized assessment allows the collection of potentially informative response latencies. It can also allow adaptive testing, where success or failure on a particular item determines whether an easier or harder item is presented next. Computerized assessment can also have administrative advantages, allowing a tester to administer assessments to a number of participants at one time, resulting in time and cost reductions (Chapelle, 2008). The Revised Token Test (McNeil & Prescott, 1978) is one example of an assessment used clinically for which a computerized version has been successfully developed for people with aphasia (PWA; McNeil et al., 2008).
Although technology may support alternative assessment methods, there are a number of questions that need to be answered before decisions can be made about a major implementation of computerized assessments in the clinical setting. For example, does test medium affect performance? And are there differences between individuals on the equivalence of these test conditions? The purpose of this study was to answer these questions and to determine whether scores obtained from PWA on a computerized version of two language assessments were equivalent to those gained by using the traditional pen-and-paper version.

There is a large body of research on neurologically normal populations that has identified issues that should be considered when investigating computerized testing, including the equivalence of scores across methods of administration, variables affecting performance on computerized assessments, and individuals’ perceptions of computerized testing.

Despite large differences between methods of testing observed in the early 1990s (Dillon, 1992), a more recent review of testing methods concluded that computerized testing is becoming more comparable to the traditional method, possibly due to advances in computer technology and/or increased familiarity of participants with computers (Noyes & Garland, 2008). Scores on traditional and computerized versions of assessments have been found to be well-correlated in a number of studies, suggesting that they are measuring the same construct, such as Raven’s Standard Progressive Matrices (RSPM; Raven, Raven, & Court, 2000; Williams & McCord, 2006), adult literacy (Chen, White, McCloskey, Soroui, & Chun, 2011), English proficiency tests (Choi, Kim, & Boo, 2003), patient outcome measures (Gwaltney, Shields, & Shiffman, 2008), and mental health assessments (Wijndaele et al., 2007). However, investigations into the differences between correlated scores, which would indicate the relative difficulty of the test medium, have produced mixed results. For example, Gwaltney et al. (2008) reported lower scores on the computerized version of their patient outcome measures compared to the pen-and-paper version, whereas Williams and McCord (2006) found no significant difference between computer-administered and experimenter-administered versions of the RSPM. These differences may be due to the task involved: Patient outcome measures, for example, are typically assessed using self-report questionnaires, whereas the RSPM is a standardized test of nonverbal ability. Some outcomes have been mixed even when using the same test. For example, there was no observed difference between scores on the Benton Visual Retention Test (Merten, 1999) when using a between-subjects design, but participants scored more poorly on the computerized version when a within-subjects design was used (Thompson, Ennis, Coffin, & Farman, 2007). This difference is possibly due to greater sensitivity in the latter or the influence of other factors such as computer experience that might vary systematically between groups of participants.

The small disadvantage sometimes found for scores on computerized tests could be due to a range of factors that may be relevant for individuals with cognitive and/or communication difficulties. Although a number of factors, including experience with and attitudes toward computers, have been investigated, the studies have provided largely contradictory findings. Computer anxiety has been linked to more extreme scores on assessments of negative affect (Schulenberg & Yuterzenka, 1999) and may have contributed to difficulties in learning and applying new rules on a computerized version of the RSPM (Kubinger, Formann, & Farkas, 1991). However, other studies have found no such effects (Thompson et al., 2007; Williams & McCord, 2006), but they did show a lack of variability in the computer anxiety scores, which may have reduced the ability to detect an effect.

The findings for the effect of ability are also inconsistent, but the differences observed are between neurologically normal and neurologically impaired research participants. Clariana and Wallace (2002), for example, found that language students with higher ability performed better on a computer-based assessment than on the paper-based version. However, in a study where PWA were assessed remotely by a clinician using a webcam and stimuli were delivered using the computer, ability as indicated by severity of language difficulties did not affect the comparability of the majority of remote assessment results to results obtained by the clinician in face-to-face assessment (Hill, Theodoros, Russell, Ward, & Wootton, 2009). Lastly, although it might seem logical to assume that a lack of computer experience would result in poorer scores on computerized tests, the evidence suggests otherwise (Clariana & Wallace, 2002).

The final issue in evaluating the equivalence of computerized and pen-and-paper tests is the participants’ perceptions of the tests. Many recent studies have found positive attitudes toward computerized testing, with more participants preferring computerized tests over pen-and-paper versions (e.g., Weber et al., 2003; Wijndaele et al., 2007), though age appears to be a factor here, with older people enjoying computerized testing less than pen-and-paper tests (Ivnik, Malec, Tangalos, & Crook, 1996). Evidence suggests that PWA react positively to computerized testing and treatment. For example, participants interviewed by Wade, Mortley, and Enderby (2003) experienced gains in confidence and reported valuing the control and independence that computer-delivered treatment gave them. PWA have also reported being satisfied by the assessment process when tested remotely using a computer and webcam (Hill et al., 2009), though in this case, there was a clinician in the room with them.

The presence versus absence of a tester in the room may affect a client’s performance and/or experience of the assessment, and yet to our knowledge, this factor has not been considered in any previous research. Indeed, often it is not clear in reported research on computerized tests whether or not the researcher was present in the room during testing. This issue requires investigation because for some people, the possibility of performing poorly in front of another person may itself inhibit their performance; alternatively, individuals may feel supported by the presence of another person (e.g., losing confidence when left alone with the computer; Weber et al., 2003).
Despite considerable attention to computerized testing of neurologically normal populations, relatively little research has considered the particular needs of PWA, and they are often excluded from studies (see, for example, Yip & Man, 2009). This is of concern because the advantages and disadvantages of computerized assessment may apply differently to PWA than they do to other populations. For example, PWA may be older and less familiar with computers than the participants in the studies reported earlier, which could affect their performance on computerized assessments. Language difficulties may make computerized assessment inherently more difficult as presentation of instructions cannot be tailored to the needs of the individual.

Aphasia often occurs alongside physical difficulties such as hemiplegia and sensory difficulties such as homonymous hemianopia, which may affect the way PWA can interact with computers. Additionally, the increased cognitive load required to operate a computer alongside performing a task reported by some research participants (e.g., Noyes, Garland, & Robbins, 2004) may be especially problematic for PWA, who may have slowed information processing or difficulties with attention (Gerritsen, Berg, Deelman, Visser-Keizer, & Meyboom-De Jong, 2003).

Our study compared computerized and traditionally administered versions of tasks with PWA. It aimed to address the equivalence of scores, the variables that affect equivalence, the efficiency of computerized testing in terms of time, and participants’ views on computerized tests. In addition to traditional pen-and-paper and fully computerized versions of assessments, our study included a computerized condition where the clinician remained in the room.

### Method

#### Participants

Fifteen PWA were recruited through a community clinic for people with acquired communication difficulties in southeast England (see Table 1 for participant details). By clinical criteria, three of the participants would be considered recovered as their aphasia quotient on the Western Aphasia Battery (WAB; Kertesz, 1982) >93.8. However, although problems at the single-word and sentence level were not marked for these three participants, their difficulties were clearly evident when abstract or less frequent words were required and in connected speech (e.g., storytelling). Eight of the participants were male and seven were female. The mean age was 59 (range 39–78). The participants were at least 12 months post stroke onset (average 80 months, range 13–225), were medically stable, had sufficient sensory abilities to identify pictures on a computer screen and hear speech in a quiet room, and did not have significant cognitive difficulties (as assessed by scores on Raven’s Coloured Progressive Matrices; Raven, Court, & Raven, 1995). Each participant had English as his or her primary language and was a past or present attendee at the clinic for group or individual speech-language pathology treatment. The participants were asked to estimate their current computer usage; this is given in hours per week. Also provided in Table 1 are details on the ethnicity, number of years of education, and previous occupation of the participants.

### Conditions

There were three testing conditions, which all participants completed. In order to minimize practice effects, each

### Table 1. Characteristics of the study participants.

<table>
<thead>
<tr>
<th>Part</th>
<th>Sex</th>
<th>Age</th>
<th>MPO</th>
<th>Race</th>
<th>Years of education</th>
<th>Previous occupation</th>
<th>RCPM</th>
<th>Current computer use (hr per wk)</th>
<th>AQ</th>
<th>Aud Comp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>66</td>
<td>225</td>
<td>W</td>
<td>20</td>
<td>School principal</td>
<td>36</td>
<td>40.00</td>
<td>65.6</td>
<td>8.50</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>47</td>
<td>44</td>
<td>W</td>
<td>15</td>
<td>Librarian</td>
<td>35</td>
<td>28.00</td>
<td>87.8</td>
<td>9.20</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>51</td>
<td>32</td>
<td>W</td>
<td>19</td>
<td>Librarian</td>
<td>35</td>
<td>21.00</td>
<td>54.6</td>
<td>9.40</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>57</td>
<td>65</td>
<td>W</td>
<td>16</td>
<td>Company director</td>
<td>27</td>
<td>15.00</td>
<td>95.0</td>
<td>9.20</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>64</td>
<td>36</td>
<td>B</td>
<td>17</td>
<td>Microbiologist</td>
<td>29</td>
<td>14.00</td>
<td>87.3</td>
<td>8.55</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>41</td>
<td>40</td>
<td>W</td>
<td>16</td>
<td>Graphic designer</td>
<td>33</td>
<td>14.00</td>
<td>92.7</td>
<td>9.85</td>
</tr>
<tr>
<td>7</td>
<td>F</td>
<td>64</td>
<td>73</td>
<td>W</td>
<td>17</td>
<td>Attorney</td>
<td>28</td>
<td>10.00</td>
<td>91.6</td>
<td>10.00</td>
</tr>
<tr>
<td>8</td>
<td>M</td>
<td>53</td>
<td>88</td>
<td>W</td>
<td>16</td>
<td>Futures analyst</td>
<td>36</td>
<td>7.00</td>
<td>61.6</td>
<td>5.70</td>
</tr>
<tr>
<td>9</td>
<td>M</td>
<td>53</td>
<td>58</td>
<td>W</td>
<td>16</td>
<td>Actor</td>
<td>31</td>
<td>6.00</td>
<td>84.4</td>
<td>10.00</td>
</tr>
<tr>
<td>10</td>
<td>F</td>
<td>76</td>
<td>13</td>
<td>W</td>
<td>13</td>
<td>Realtor</td>
<td>32</td>
<td>4.00</td>
<td>94.3</td>
<td>9.75</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>60</td>
<td>52</td>
<td>W</td>
<td>18</td>
<td>Attorney</td>
<td>27</td>
<td>4.00</td>
<td>73.5</td>
<td>7.85</td>
</tr>
<tr>
<td>12</td>
<td>F</td>
<td>39</td>
<td>96</td>
<td>B</td>
<td>13</td>
<td>Nurse</td>
<td>33</td>
<td>2.00</td>
<td>88.6</td>
<td>9.30</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>75</td>
<td>108</td>
<td>W</td>
<td>10</td>
<td>Labor union leader</td>
<td>28</td>
<td>2.00</td>
<td>71.4</td>
<td>7.70</td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>78</td>
<td>218</td>
<td>W</td>
<td>16</td>
<td>Teacher</td>
<td>34</td>
<td>0.75</td>
<td>98.0</td>
<td>10.00</td>
</tr>
<tr>
<td>15</td>
<td>F</td>
<td>61</td>
<td>48</td>
<td>W</td>
<td>16</td>
<td>Receptionist</td>
<td>35</td>
<td>0.00</td>
<td>93.2</td>
<td>9.60</td>
</tr>
</tbody>
</table>

**Note.** Part = participant; MPO = months post onset of stroke; RCPM = Raven’s Coloured Progressive Matrices (Raven, Court, & Raven, 1995); WAB = Western Aphasia Battery (Kertesz, 1982); AQ = aphasia quotient; Aud Comp = Auditory Sentence Comprehension subtest; M = male; F = female; W = White; B = Black.

*a*Children start school at the age of 5 years in the United Kingdom. *b*None of the participants in the study was currently employed.
condition involved two tasks (described below) but with different items. Items were matched across conditions to ensure compatible levels of difficulty. Care was taken to ensure that instructions given in the computerized version of the tasks were as similar as possible to those given in the pen-and-paper versions (e.g., the same wording was used).

**Computer-only condition.** The participant was alone in the room and all stimuli and instructions were presented by the computer. Written and illustrated instructions were presented on the screen (see Figure 1 for a screen shot), and an audio recording of the instructions was played.

**Computer and clinician condition.** The participant was alone in the room throughout the tasks, read the instructions to the participant (while they were on the screen), and answered questions and gave general feedback (i.e., minimal encouraging comments to maintain the participants' interest) during the tests. The clinician sat out of the sight line of the participant while the tasks were completed and did not engage in any additional activity during this time.

**Pen-and-paper condition.** The assessments were administered as directed in the published versions. The instructions were read by the clinician with the visual prompts (e.g., an example array of four pictures for the sentence-picture matching task) but not the written text. The clinician gave feedback and help similar to the computer and clinician condition. There were two versions of the pen-and-paper test, with different items, which all participants completed. Two versions of this condition were included in order to estimate test–retest reliability and to provide a point of comparison for interpreting differences between the computerized and pen-and-paper conditions. In order to do this, one of the pen-and-paper conditions was always presented last.

The start and end times of each condition were recorded by the researcher.

**Tasks**

Two tasks of language comprehension were selected to compare across conditions: sentence-picture matching and grammaticality judgment. Four versions of each task were constructed (in the case of sentence-picture matching, by modifying the published version to use each of the distractor pictures) to form the testing conditions.

**Sentence-picture matching.** Participants saw four pictures and heard a sentence that was read aloud. Participants were asked to point to the picture, from an array of four pictures, that best matched the sentence. Stimuli were taken from the Auditory Sentence Comprehension subtest in the Comprehensive Aphasia Test (17 items, CAT; Swinburn, Porter, & Howard, 2004) and from the Test for Reception of Grammar (13 items, TROG; Bishop, 2003) to form a set of 30 items. The visual stimuli were identical across the conditions, and each spatial position in the array was correct an equal number of times in each condition. In order to minimize the effects of learning, the order of the items was randomized differently within each condition. As the items were randomized rather than being in order of difficulty, there was no discontinue rule. Four practice items were selected from the TROG and were presented at the start of the assessment, and in this practice task, participants were shown the correct answer after they made their response. Participants’ responses to practice items were not included in the data analysis. The instructions from the CAT were used for the test.

**Grammaticality judgment.** Participants were asked to say whether each sentence was good or bad by pointing to a tick or a cross on the screen. The stimuli for this task were 184 items from the set used by McDonald (2000). The sentences used a range of grammatical structures or elements that were omitted or changed in the ungrammatical versions (e.g., “A shoe salesman sees many feelfeets throughout the day”; “The girl is writingwrite a letter to her mother”). The sentences were divided between the four conditions such that the grammatical and ungrammatical versions of a sentence were in different conditions. Each condition had equal numbers of grammatical and ungrammatical items, and the types of structure (past tense, pronouns, etc.) were distributed as evenly as possible among the conditions, with 46 sentences in each condition. The order of items was randomized differently within each condition. Four practice items were presented at the start of each condition, from Seol (2005), and again participants were shown the correct answer after they had made their response. As with the sentence-picture matching task, practice items were not included in the data analysis. Instructions were taken from the grammaticality judgment task in the Verb and Sentence Test (Bastiaanse, Edwards, & Rispens, 2002).

**Computerized Tasks**

The computerized tasks were administered using a desktop PC in a quiet but not soundproofed room.
Responses were collected using a touchscreen interface (Keytec 17” Touch Screen KTMT-1700). The tasks were delivered by a customized computer program written in Visual Basic .NET. Visual stimuli were displayed on the computer screen, and auditory stimuli and instructions were played at the appropriate points in the tasks via the PC’s internal speakers. Recordings of stimuli and instructions used in the computerized versions of the tasks were made by the same clinician (a speech-language pathologist) who administered the tasks in the pen-and-paper conditions. The computerized tasks were controlled by the participant using Repeat and Continue buttons that remained on the screen at all times (see Figure 2 for a screen shot from the grammaticality judgment task). Participants could repeat the instructions, the practice sessions, and the auditory stimuli (one repeat only).

In the practice sessions, feedback was given by a red box around the correct response, which was shown for 2 s after the participant pressed continue. The computer gave no feedback during the main tasks. Participants were allowed to self-correct their responses by making a second response. The first response was taken as their answer unless there was a subsequent different answer. Participants were also given an opportunity to practice using the touchscreen before the experiment began in order to minimize the risk of their responses not being recorded by the computer. Practice involved using the touchscreen to move the mouse pointer around on the screen before testing, to work through the pages of instructions for the tasks, and to carry out the practice tasks.

The computerized versions of the tasks were constructed to be as similar as possible to the pen-and-paper versions and to be accessible to people with language difficulties. In the sentence-picture matching task, the pictures were shown on the screen for 5 s and then the pictures disappeared and the sentence was played. When the sentence ended, the pictures reappeared and the participants could select their answer by touching the correct picture. In the grammaticality judgment task, the sentence was played and then a tick and cross were shown on the screen for the participant to touch to give their answer. In both tasks, the participant had to press continue to progress to the next item.

### Questionnaires

Three questionnaires were used to collect information about the participants’ computer use (before the study) and their experiences of the testing methods (subsequent to testing). Questionnaires were administered by the clinician conducting the tasks, though this person was unknown to the participants before testing and had no involvement in their subsequent management. When a questionnaire item expressed a negative statement, to ensure that participants were giving responses that reflected their views, the clinician paraphrased the question to reflect the answer they had given (e.g., “You said three and that means you disagree and so you do understand how to use software”).

Before testing commenced, participants completed the computer aversion items from the Computer Aversion, Attitudes, and Familiarity Index (CAAFI; Schubenberg & Melton, 2008), which provides a 10-item measure of participants’ degree of aversion to computers. Participants indicated their response to statements such as “When I use a computer, I am afraid that I will damage it” and “I am smart enough to use a computer” using a visual 7-point Likert scale (ranging from absolutely false to absolutely true). The minimum possible score on this part of the CAAFI is –30, reflecting extreme aversion; the maximum is 30.

After each testing condition, the participants completed a questionnaire on their experience of that condition. The questionnaire was specifically constructed for this study and asked questions about the quality of the stimuli, the ease of response, and whether the participant enjoyed the condition. Participants indicated their response using a 7-point Likert scale (ranging from not at all to very much). The experience questionnaire is provided in the Appendix.

At the end of the final condition, participants were asked to complete a method preference questionnaire that was adapted from the one used by Thompson et al. (2007). This questionnaire asked the participants to select their favorite and least favorite conditions from picture symbols representing four possible responses (pen-and-paper task, computer and clinician, computer only, no preference). Although there are limitations with this approach (e.g., individuals may use different criteria to judge their favorite), its aim was to elicit participants’ subjective feelings about the testing conditions.

### Procedure

The study used a within-subjects design, with each participant completing all three conditions, plus a second pen-and-paper condition. To reduce the impact of practice effects, the order of conditions was counterbalanced.

---

**Figure 2.** Screen shot showing computerized delivery of the grammaticality judgment task.
resulting in six unique orders of the first three conditions (ABC, ACB, BAC, BCA, CAB, CBA); one of the pen-and-paper tasks was always last. Each order of conditions was completed by at least two participants. The order of tasks was consistent across the conditions and participants so that the sentence-picture matching task was always first and the grammaticality judgment task was always second. Before the participants started the tasks, they were reminded to respond as quickly and accurately as they could. The testing was run over two 90-min testing sessions at least 2 weeks apart, with two conditions in the first session and two in the second. Before testing, the participants read an information sheet, completed a consent form, and completed the computer use questionnaire with the clinician. Participants also completed the experience questionnaire after each condition and the method preference questionnaire at the end of the final condition.

**Results**

**Equivalence Across Methods of Administration**

Two parameters are important when judging equivalence between modalities. First, the correlation between scores from the two versions of the tasks is an indication of whether the tasks are sensitive to the same factors and whether they are measuring the same construct. The second parameter is whether there is a significant difference between the scores from the computerized and pen-and-paper tasks, which indicates the relative difficulty (see, for example, Mead & Drasgow, 1993; Williams & McCord, 2006).

The scores from the pen-and-paper condition were significantly strongly correlated with those from the other pen-and-paper condition, the computer-only condition, and the computer and clinician condition for both tasks (see Table 2).

Means for the two pen-and-paper conditions were compared for the two tasks by two paired-samples *t*-tests. There was no significant difference between the first pen-and-paper condition (*M* = 24.40, *SD* = 4.03) and the second pen-and-paper condition (*M* = 23.27, *SD* = 7.95) in the sentence-picture matching task, *t*(14) = .658, *p* = .521; there was also no significant difference between the first pen-and-paper condition (*M* = 40.00, *SD* = 4.24) and the second pen-and-paper condition (*M* = 37.53, *SD* = 11.15) in the grammaticality judgment task, *t*(14) = .857, *p* = .406.

To examine differences between these mean scores for the two tasks, one-factor repeated measures analyses of variance (ANOVA) were performed (Mauchly’s test indicated that the assumption of sphericity was not violated in either case). Only the scores from the first version of the pen-and-paper condition were used in order to minimize the effect of practice on scores and because only this version was counterbalanced with the other conditions (see Procedure above). There was a significant effect of condition on scores in the sentence-picture matching task, *F*(2, 26) = 11.912, *p* < .001, *η*² = .478, and post hoc pairwise Bonferroni-corrected comparisons (with a corrected *p* value of .008) revealed that scores from the computer-only condition were lower than those from both the pen-and-paper (*p* = .002) and the computer and clinician condition (*p* = 0.08), though the latter did not reach statistical significance. There was also a significant main effect of condition on scores in the grammaticality judgment task, *F*(2, 26) = 7.91, *p* = .002, *η*² = .378, with scores in the computer and clinician condition being lower than those in the pen-and-paper condition (*p* = .004). Figure 3 shows the mean scores in the three conditions that were compared in the ANOVAs.

**Variables Affecting Performance on Computerized Assessments**

We examined the use of explanatory variables to predict which participants showed differences between computerized and pen-and-paper task scores. Explanatory variables investigated were age, months since stroke, WAB aphasia quotient, auditory comprehension, computer aversion score (shown in Figure 4), and hours of computer use per week. Difference scores were computed between the pen-and-paper task scores and the computerized task scores, and the relationship between the difference scores and possible explanatory variables was explored using Pearson’s product–moment correlation coefficients. None of the possible explanatory variables correlated significantly with any of the difference scores.

**Task length.** Computerized tasks are frequently assumed to be a more time-efficient way of administering assessments, but this has seldom been tested. In the present study, the number of minutes to administer both tasks in each condition was recorded, and this was compared between conditions using a one-factor repeated measures ANOVA. As above, only the score from the first version of the pen-and-paper condition was used in order to minimize the effect of practice on scores. Mauchly’s test indicated that the assumption of sphericity was violated, *χ*²(2) = 8.301, *p* = .016, so the Greenhouse-Geisser correction was used. There was no main effect of condition on time taken to complete the tasks, *F*(1,359, 19.02228) = .648, *p* = .477, *η*² = .044, and no significant difference in the time taken between the three conditions. Figure 5 shows the time taken to complete each condition.

| Table 2. Pearson’s product–moment correlation coefficients for the relationships between one pen-and-paper condition and the other pen-and-paper condition, the computer-only condition, and the computer and clinician condition for both the sentence-picture matching task and the grammaticality judgment task. |
|---|---|---|
| | Pen and paper | Computer only | Computer and clinician |
| Sentence-picture matching | .89** | .95** | .92** |
| Grammaticality judgment | .69** | .87** | .78** |

*p* < .01; **p* < .001.
Figure 3. Mean scores for the three conditions for the two tasks (error bars show 95% confidence intervals; asterisks indicate statistically significant differences).

Figure 4. Participants’ computer aversion scores from the Computer Aversion, Attitudes, and Familiarity Index (Schulenberg & Melton, 2008) questionnaire.
Perceptions of computerized testing. In a questionnaire following each condition, participants were asked to rate different aspects of the experience on a 7-point Likert scale. These ratings were compared among conditions, with only the ratings from the first pen-and-paper condition presented included in the analysis (as above). As the data were not continuous, a Wilcoxon signed-rank test was used to compare the paired samples (a total of 33 comparisons). The average ratings are shown in Table 3.

Participants rated the ease of giving their responses significantly more highly in the pen-and-paper condition than in the computer-only condition, $Z = -2.95, p = .003, r = - .44$, or in the computer and clinician condition, $Z = -2.40, p = .016, r = -.36$. Although the response method was identical, participants also found it easier to give their responses in the computer and clinician condition compared to the computer-only condition, $Z = -2.22, p = .027, r = -.33$. Participants’ ratings of the clarity of the instructions were higher for the pen-and-paper condition than for the computer-only condition, $Z = -2.23, p = .026, r = -.33$, despite identical wording, with the only difference being that in the computer condition, the participant also saw written instructions. Participants rated their enjoyment of the tests as higher in the computer and clinician condition than in the computer-only condition, $Z = -2.06, p = .04, r = -.31$, but there were no other significant differences. There were no significant differences between conditions in participants’ other ratings.

Results from the method preference questionnaire indicated that approximately half of the participants expressed no preference for testing condition when considering different aspects of the experience (see Figure 6). However, when participants expressed a preference, the pen-and-paper version was the preferred method, and the computer-only version was the least popular. Participants’ overall preference for different conditions (i.e., best overall vs. worst overall) was examined using chi-square analysis to explore whether favorite and least favorite choices were distributed differently between the conditions. Although there was no significant pattern of preference between the computer and clinician condition and either the computer-only or the pen-and-paper condition, the pen-and-paper condition was significantly preferred over the computer-only condition, $\chi^2(1) = 4.81, p = .028$.

When participants completed the method preference questionnaire and indicated which methods they considered favorite and least favorite, they were asked to comment on why they made those choices. These responses were transcribed by the clinician conducting the experiment. Six participants commented that they enjoyed the interaction with the clinician, making statements such as “I like human contact … looking at the person in front of me … the computer is impersonal.”

Another six participants said that they found the clinician’s presence reassuring, particularly in the computer and clinician condition. These statements included “You’ve always got a little bit of help in case something goes wrong” and “Computers don’t always work right … If you’re on your own you don’t even know if you’re doing it right.”

When participants made comments about the use of the computer in carrying out the assessments, six participants expressed anxiety or uncertainty about using the computer, saying that they didn’t know if they were doing the tests properly or that they were scared of getting it wrong. For example, “I don’t have a computer … scared to repeat or spoil sometimes … never-ending … makes me feel nervous because I don’t normally use them.”

However, other main ideas about the computer were positive. Three participants said that they liked or felt confident using the computer, for example, “I’m used to the computer … [prompt – why does that make it better?] … I’m in control, I understand it.” Two participants also commented that they liked that the computer program allowed them to work at their own pace, for example, “It’s my time alone. I can do it fast or slow, you know. When I’m working I just want free time.”
Discussion

We compared the use of computerized and traditionally administered tasks to assess PWA. The conditions were compared in terms of the equivalence of the resulting scores, which variables predicted the differences between conditions, the time taken to administer each condition, and participants’ experience of the assessments.

The study demonstrated a strong correlation between scores from computerized and pen-and-paper tasks, which has not previously been demonstrated in PWA. This finding is consistent with previous reports concerning participants with no cognitive or communication difficulties (e.g., Williams & McCord, 2006). The strong correlation between the scores from the pen-and-paper and computerized versions of tasks suggests that the computerized format is sensitive to the same factors as the traditionally administered tasks. This result demonstrates that computerized tasks could be used to assess PWA, though caution should be exercised because lower scores were observed in the conditions that involved the computer (see Figure 3).

There were no significant differences between scores on the two versions of the pen-and-paper conditions, which suggests that there was consistency of performance over time, which provides some confidence in our interpretation of differences between the pen-and-paper and computerized conditions. The lower average scores from the computerized versions of the language tasks are consistent with the results of some previous studies (e.g., Chen et al., 2011). This difference has been attributed to various factors, including an increased cognitive load in computerized tasks (Wastlund, Reinikka, Norlander, & Archer, 2005); higher cognitive load may be one of the factors underlying the lower scores on the computerized tasks in this study. Although increased difficulty of the computerized tasks was not a main theme in participants’ comments, one person observed,

When I was using the computer I worried about whether it was right or not and then what you’re trying to do gets scrambled. So you’re thinking about how you’re doing what you’re doing rather than what you’re doing.

Knowing that computerized tasks may lead to lower scores is important when considering whether to use them clinically. The difference means that scores from computerized versions of tasks should not be compared to scores from the same tasks administered in traditional formats, and neither should they be compared to norms collected using the pen-and-paper version. If pretreatment baseline scores were collected via computer, then the same method should be used when collecting posttreatment measures.
The findings of this study suggest that some people perform better in the presence of a clinician for some language tasks, even if that person is not actively involved in the assessment process. However, not all tasks are equal in this respect, and our study showed a different pattern of results across the two tasks (see also Schroeders & Wilhelm, 2011). Participants expressed relatively neutral views on this condition: The addition of the clinician in the computerized task was only significantly better rated on ease of response and whether the participants had enjoyed doing the tasks that way. In the method preference questionnaire, the computer and clinician condition showed no significant pattern of preference and was rarely selected as the least or most favorite condition. However, participants’ qualitative comments indicated that they preferred doing the tasks with the researcher in the room, with one participant describing that person as being like a “lifeguard.”

This study did not find any difference in the average time taken for each condition, though of course in the computer-only condition, the clinician was able to leave the room. This could be advantageous in a clinical setting. Although computerized testing may not be quicker for the individual client, it may be more time-efficient for the clinician, freeing him or her for planning or administration tasks. With preparation, the clinician would also be able to use this time even if the client preferred that he or she remain in the room. However, there may be disadvantages to using computerized tasks without the clinician being present. For example, the lack of opportunity for the clinician to make observations during assessment would mean that any information that the computer does not pick up (such as relative difficulty of stimuli and nonverbal cues) would not be recorded. Furthermore, although favored by some, the computer-only condition was not popular among many of the participants in our study. Although comparisons of the experience ratings that participants gave for each condition were consistently high across the testing conditions (M > 5 on the 7-point rating scale), the method preference questionnaire revealed a marked preference for the pen-and-paper administration of the tasks. This mirrors the findings of Ritchie and Newby (1989), who used a similar study design and reported that college students rated instruction via television less enjoyable than when the instructor was in the room with them.

A negative experience of the assessment process may have a detrimental effect on the relationship between a client and his or her clinician and the client’s engagement with treatment. It is possible that this effect may be mitigated by incorporating—as some treatment programs have done—an avatar whose role is to guide the user through tasks (e.g., Lee & Cherney, 2008). However, computerized assessment may be more suitable for some clients than others. It was clear from the subjective ratings and comments that some participants enjoyed using the computer and felt that it allowed them to work at their own pace (though these attitudes were not reflected in higher—or lower—scores on the tasks on the computer).

This study identified no relationships between potential explanatory variables and the difference between scores on computerized and pen-and-paper versions of tasks. The failure to find effects in this study might be attributed to the low number of participants or to high variability in participants’ scores on the tasks. However, it should be noted that previous studies that have found an effect of variables such as computer anxiety have tended to be tests of personality and mood (Schulenberg & Yutzy, 1999), whereas the studies that have found no effects have used tasks more similar to the present study, such as ability tests and psychological tests (Thompson et al., 2007; Williams & McCord, 2006). If computerized testing were to be used in the clinic, it would be useful to be able to identify participants who would most likely to score poorly on computerized tests so that they could be selected for pen-and-paper testing. However, none of the variables investigated in the present study would allow this, and further research would therefore be required.

A note of caution is of course required in the interpretation of the findings of this relatively small-scale study. Only two language tasks were included; results observed may not be found in a comparison of these conditions with different types of assessment. Additionally, although every effort was made to ensure comparable levels of difficulty across the conditions by matching the sets of stimuli, it was not practical in a study of this size to counterbalance the items in each condition, and this may have resulted in unintended differences in level of difficulty between the pen-and-paper and computerized conditions. Participants were questioned about their experiences of the testing methods by the clinicians who administered them, which may have affected their rating and reported preferences, though the clinician herself was careful not to indicate a preference.

As mentioned above, it may be that the low number of participants prevented some significant findings from being observed, though this may increase confidence in the robustness of the significant differences that were found. In addition, although they represent a range of ages and levels of experience with and aversion to computers, the participant group was drawn from a relatively small area in the southeast of England and was almost all White, and all but three participants were educated past the age of 18. As such, their responses—and views—may not necessarily be representative of all PWA. However, computerized testing might have been expected to receive a more positive response from highly educated adults, many of whom had used computers on a daily basis before their stroke, even though their use decreased post stroke.

The use of computers in speech-language pathology is likely to increase in the future, and so further research in this area is warranted with a larger group of participants and with a range of different assessments. Assessments delivered by computers offer well-documented advantages over pen-and-paper assessments. As well as the potential to free up clinician time, it is possible with computerized tasks to make testing more efficient by adapting to the performance of
individuals (e.g., discarding items that are too easy), thus determining an individual’s ability with a shorter test. However, it is clear that, as with tests that have been used with neurologically normal populations, caution must be exercised over the use of computers in assessment. Clinicians should critically evaluate computerized tests; should be responsible for ensuring that they are appropriately trained in administering computerized tests; and should be aware of the psychometric properties of the test, that is, whether the results can be compared to standard norms (Schulenberg & Yutrzenka, 2004). One proposed alternative to single computerized tests is to view computer-delivered testing as an ongoing form of assessment to monitor progress (Petheram, 2004). This avoids many of the disadvantages of computerized testing by comparing the participant to his or her own previous performance and can form a part of outcome measurement in conjunction with tests of generalization to different items and situations. Although there is little doubt that computers will feature large in the clinic of the future, on judging whether a computer version of a test is appropriate, the equivalence of the test to the paper-based version may depend on the construct or ability being tested. The suitability of the test will depend on the preferences of the individual being tested.

Acknowledgment

The authors wish to thank the participants who contributed to the study.

References


Appendix

Experience Questionnaire Following Each Test Condition

Think about the test you have just done.

Stimuli
1. Was the sound good quality?
   - Not at all
   - Very much

2. Were the pictures good quality?
   - Not at all
   - Very much

3. Were the questions a fair way to test your understanding?
   - Not at all
   - Very much

Comfort
4. Were you physically comfortable during the test?
   - Not at all
   - Very much

5. Did you find this an easy way to make responses?
   - Not at all
   - Very much

Test
6. Were the instructions easy to understand?
   - Not at all
   - Very much

7. Was this kind of test (on the computer/with me asking questions) a fair way to test your understanding?
   - Not at all
   - Very much

8. Did this way of testing make you nervous? (intimidated)
   - Not at all
   - Very much

9. Are you confident in the results of this test?
   - Not at all
   - Very much

Overall satisfaction
10. Did you enjoy the tests (on your own at the computer/on the computer with me still here/with me asking you questions)?
    - Not at all
    - Very much

11. Would you be happy to do this type of test again?
    - Not at all
    - Very much