A COMPARISON OF COMPUTER-DELIVERED AND PAPER-BASED LANGUAGE TESTS WITH ADULTS WHO HAVE APHASIA

Caroline Newton*

Developmental Science, Division of Psychology and Language Sciences,

University College London

Kadia Acres

Adult Speech and Language Therapy, The Dudley Group NHS Foundation Trust

Carolyn Bruce

Language and Communication, Division of Psychology and Language Sciences,

University College London

* Corresponding author

Dr C. Newton, UCL Developmental Science, Chandler House, 2 Wakefield Street, London, United Kingdom, WC1N 1PF

Tel: +44(0)20 7679 4222

E-mail: caroline.newton@ucl.ac.uk

Abstract

Purpose: This study investigated whether computers are a useful tool in the assessment of adults with aphasia. In order to do this, computerized and traditionally-administered tests were compared to determine whether i) the scores were equivalent, ii) the administration was comparable, and iii) the participants' perceptions of the different assessment methods were similar.

Method: Fifteen participants with aphasia were assessed on two language tasks – sentence-picture matching and grammaticality judgment – in three conditions: computer only, computer with the researcher present, and using standard methods. The participants also completed questionnaires rating aspects of each condition.

Results: Scores from the traditionally-administered tests were highly correlated with those from computerized tests, but scores from computerized tests were significantly lower. While some individuals felt comfortable with the computer, overall participants preferred the traditional assessment method or where another person was in the room. No factors were identified that predicted performance in the computer condition.

Conclusions: The results suggest that people with aphasia can be assessed using computerized tests, but that caution should be exercised when comparing scores to those collected using traditional methods, including norms. The variation in opinion regarding the computerized tests suggests that this method might be more suitable for some participants than others.

Introduction

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. Computers are becoming increasingly used particularly in the treatment of aphasia (e.g., Ramsberger & Marie, 2007; Palmer, Enderby, & Hawley, 2007; Mortley, Wade, Enderby, & Hughes, 2004; Wertz & Katz, 2004). In a survey of US speech-language pathologists (SLPs), Davis and Copeland (2006) found that more than 50% of respondents used computers in aphasia treatment, including treating reading, writing and word-finding difficulties. Assessing aphasia using computers is less common, though with increasing pressures on SLP services in terms of both staff numbers and staff time, computer-delivered assessment may have great potential to relieve some of those pressures and bring both financial and administrative advantages. For example, the use of videoconferencing 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 their own responses into the computer either in the clinic or online, including the fact that this could be carried out without an SLP present. These 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 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. It can also have

organizational advantages, allowing a tester to administer assessments to a number of participants at one time, resulting in time and cost reductions (Chapelle, 2008). Indeed, the Revised Token Test (McNeil & Preston, 1978) is one example of an assessment used clinically for which a computerized version has been successfully developed with people with aphasia (McNeil et al., 2008). Although technology may support alternative assessments, there are a number of questions that need to be answered before decisions are 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 individuals with aphasia on a computerized version of two language assessments are equivalent to those gained by using the traditional pen and paper format.

There is a large body of research on neurologically normal populations which has identified issues that should be considered in investigating this process: equivalence of scores across methods of administration, variables affecting performance on computerized assessments and individuals' perceptions of computerized testing.

Despite worryingly large differences between modalities observed in the early 1990s (Dillon, 1992), a more recent review concludes that, though computer performance is generally poorer than pen and paper, 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 have been found to be well-correlated in a number of studies suggesting that they are measuring the same construct, e.g., Raven's Standard Progressive Matrices (RSPM; Williams & McCord, 2006), adult literacy (Chen, White, McCloskey, Soroui, & Chun , 2011), English

proficiency tests (Choi, Kim, & Boo, 2003), patient outcome measures (Gwaltney, Sheilds ,& 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 been mixed. For example, Gwaltney et al. (2008) report lower scores on the computerized version of their patient outcome measures than the pen and paper version; Williams & McCord (2006) found no significant difference between computer administered and experimenter administered RSPM. These differences may be due to the task involved. 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 when using a between-subjects design (Merten, 1999) 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 because of 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 which may be relevant for individuals with cognitive and/or communication difficulties. Although a number of factors, including experience and attitudes towards 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 & Yutrzenka, 1999) and may have contributed to difficulties in learning and applying new rules on a computerized version of the Ravens Standard Progressive Matrices (Kubinger, Formann, & Farkas, 1991). However, other studies have found no such effects (Williams & McCord, 2006; Thompson et al., 2007), though they 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. Clariana and Wallace (2002) 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 participants with aphasia were assessed remotely by a therapist using a webcam and stimuli were delivered using the computer, severity of language difficulties did not affect the comparability of the majority of remote assessment results to results obtained by the therapist in face to face assessment (Hill, Theodorus, Russell, Ward, & Wootton, 2009). Lastly, although it might seem logical to assume that 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 pen and paper and computerized tests is the participants' perceptions of the tests. Many recent studies have found positive attitudes towards 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 people with aphasia react positively to computerized therapy and testing. For example, participants interviewed by Wade and colleagues (2003) reported valuing the control and independence that computer-delivered therapy gave them and experienced gains in confidence. People with aphasia 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 therapist in the room with them. The presence versus absence of the 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 is present in the room during testing. This issue requires investigation as 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.

Despite considerable attention to computerized testing of neurologically normal populations, relatively little research has considered the particular needs of people with aphasia 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 people with aphasia. For example, people with aphasia may be older and less familiar with computers than the participants in the studies reported above, which could affect performance. 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 people with aphasia 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 people with aphasia who may have slowed information processing or difficulties with attention (Gerritsen, Berg, Deelman, Visser-Keizer, & Meyboom-De Jong, 2003).

The study reported here compared computerized and pen and paper assessment of people with aphasia. 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, the present study included a computerized condition where the researcher remained in the room. Previous studies do not always

specify whether the researcher was present or consider that the participants might find this important (e.g., losing confidence when left alone with the computer; Weber et al., 2003).

Method

Participants

Fifteen participants with aphasia were recruited through a community clinic for people with acquired communication difficulties in South East 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) is greater than 93.8. However, whilst problems at the single-word and sentence level were not marked for these participants, their difficulties were clearly evident when connected speech was required (e.g. story telling). The mean age was 59 (range 39-78) and eight of the participants were male. Participants were at least 12 months post stroke onset (average 80 months, range 13-225), 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 the Ravens Coloured Progressive Matrices; see table 1). Each participant had English as their primary language and was a past or present attendee at the clinic for group or individual speech-language pathology. Each of the participants was asked to estimate their computer usage; this is given in hours per week. Also provided in table 1 are details on the number of years of education and occupation of the participants.

Insert table 1 about here

Conditions

There were four conditions, which all participants completed. Each condition involved both tasks (described below) but with different items, to minimize practice effects. Items were matched across conditions to ensure compatible levels of difficulty. Care was taken to ensure that instructions given in the computer-based tasks were as similar as possible to those 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 screenshot) and an audio recording of the instructions was played.
- 'Computer and Therapist' condition: the researcher remained in the room throughout the tasks, read the instructions to the participant (while they were on the screen), answered questions and gave general feedback (i.e., minimal encouraging comments to maintain the participants' interest) during the tests, as in the pen and paper condition. The researcher sat out of the eye-line of the participant while they completed the tasks, but 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 researcher with the visual prompts (e.g., an example array of four pictures for the sentence-picture matching task) but not the written text. The researcher gave similar feedback and help to the 'Computer and Therapist' 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 form a point of comparison for interpreting difference among computer-administered and pen and paper conditions.

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

Insert figure 1 about here

Tasks

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

- Sentence-picture matching

Participants were shown 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. The stimuli were taken from the auditory sentence comprehension test in the Comprehensive Aphasia Test (17 items, CAT; Swinburn et al., 2004) and from the Test of 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. The order of the items was randomized differently within each condition to minimize effects of learning. 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 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 with minor modifications for the computer versions.

- Grammaticality judgment

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 which were omitted or changed in the ungrammatical versions, e.g., "A shoe salesman sees many *feet/feets* throughout the day"; "The girl *is writing/write* 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; types of structure (past tense, pronouns, etc.) were distributed as evenly as possible between the conditions with 46 sentences in each condition. The order of items was randomized differently within each condition. Participants were asked to say whether each sentence was 'good' or 'bad' by pointing to a tick or a cross on the screen. 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) with minor modifications for the computer versions.

Computer administered tasks

The computer-based tests 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 bespoke computer program written in Visual Basic .NET. Visual stimuli were displayed on the computer screen, 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 computer versions of the tasks were made by the same researcher (a speech-language pathologist) who administered the tasks in the pen and paper conditions. The computer tasks were

controlled by the participant using 'Repeat' and 'Continue' buttons which remained on the screen at all times (see figure 2 for a screenshot 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 two seconds 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. Before the participants started the tasks, they were reminded to respond as quickly and accurately as they could. They were also given an opportunity to practice using the touchscreen before the experiment began to minimize the risk of their responses not being recorded by the computer. The participant used the touchscreen to move the mouse pointer around on the screen prior to testing and to work through the pages of instructions for the tasks and to carry out the practice tasks.

The computer versions of the tests 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 five seconds then the pictures disappeared and the sentence was played. When the sentence had finished, 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.

Insert figure 2 about here

Questionnaires

Three questionnaires were used to collect information about the participants' computer use and their experiences of the testing methods. Questionnaires were administered by the researcher conducting the tasks, though this person was unknown to the participants prior to testing and had no involvement in their subsequent management. To ensure that participants weer giving reposens that reflected their views, the researcher 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; Schulenberg & Melton, 2008), which provided 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 seven-point Likert scale (ranging from 'absolutely false to 'absolutely true'). The minimum possible score on this part of the CAFFI is -30, reflecting extreme aversion; the maximum is 30.

After each condition, 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 visual seven-point Likert scale (ranging from 'not at all' to 'very much'). The experience questionnaire is given in the appendix.

Finally, participants were asked to complete a method preference questionnaire at the end of the final condition, adapted from the one used by Thompson and colleagues (2007). The questionnaire asked participants to select their favorite and least favorite conditions from picture symbols representing four possible responses (pen and paper task, computer with therapist, computer alone, 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 conditions.

Procedure

The study used a within subjects design with each participant completing all four conditions. To reduce the impact of practice effects the order of conditions was counterbalanced, 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. The order of tasks was consistent across the conditions and participants: the sentence-picture matching task was always first; the grammaticality judgment task always second. The testing was run over two 90-minute testing sessions at least two 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 researcher. They also completed questionnaires after each condition and the method preference questionnaire at the end of the final condition (see above).

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 test is an indication of whether the tests 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, Williams & McCord, 2006; Mead & Drasgow, 1993).

The scores from the pen and paper condition were significantly strongly correlated with those from the other pen and paper condition, the computer alone condition and the computer and therapist condition for both tasks (see table 2).

Insert table 2 about here

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 ANOVAs 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. There was a significant effect of condition on scores in the sentence-picture matching task ($F(2,26)=11.912$, $p<.001$, $\eta_p^2=.478$), and post-hoc pairwise Bonferroni-corrected comparisons revealed that scores from the computer alone condition were lower than those from both the pen and paper ($p=.002$) and the computer and therapist condition ($p=0.08$; non-significant). There was also a significant main effect of condition on scores in the grammaticality judgment task ($F(2,26)=7.91$, $p=.002$, $\eta_p^2=.378$), with scores in the computer and therapist condition being lower than those in the pen and paper condition ($p=.004$). See figure 3 for the mean scores in the three conditions compared in the ANOVAs.

Insert figure 3 about here

Variables affecting performance on computerized assessments

The use of explanatory variables to predict which participants showed differences between computerized and pen and paper test scores was examined. Explanatory variables investigated were age, months since stroke, Western Aphasia Battery aphasia quotient (WAB; Kertesz, 1982), 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 test scores and the computerized test 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.

Insert figure 4 about here

Test length

Computerized tests 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 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 ($\chi^2(2)=8.301$, $p=.016$), so the Greenhouse-Geisser correction was used. There was no main effect of condition on time taken to complete the tests ($F(1.359,19.02228)=.648$, $p=.477$, $\eta_p^2= .044$), and therefore no significant difference in the time taken between the three conditions. Figure 5 shows the time taken to complete each condition.

Insert figure 5 about here

Perceptions of computerized testing

In a questionnaire following each condition, participants were asked to rate different aspects of the experience on a seven-point Likert scale. These ratings were compared between 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. The average ratings are shown in table 3.

Insert table 3 about here

Participants rated the ease of giving their response significantly more highly in the pen and paper condition than in the computer alone condition (Z=-2.95, p=.003, r=-.44) or in the computer and therapist 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 therapist than the computer alone 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 alone 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 therapist condition than in the computer 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 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 alone the least popular. Participants' overall preference for different conditions (i.e. best overall versus worst overall) was examined using chi-square analysis to explore whether favorite and least favorite choices were distributed differently between the conditions. While there was no significant pattern of preference between the computer and therapist condition and either the computer only or pen and paper conditions, pen and paper was significantly preferred over computer only by the participants ($\chi^2(1)=4.81$, $p=.028$).

Insert figure 6 about here

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 had made those choices. These responses were transcribed by the researcher conducting the experiment. Six participants commented that they enjoyed the interaction with the researcher. For example:

> *I like human contact . . . looking at the person in front of me . . . the computer is impersonal . . .*

Another six participants said that they found the researcher's presence reassuring, particularly in the computer and therapist condition. For example:

> *You've always got a little bit of help in case something goes wrong. . .*
>
> *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

This study aimed to compare the use of computerized and traditionally administered tests in the assessment of people with aphasia. 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 has demonstrated a strong correlation between scores from computerized and pen and paper tests, which has not previously been demonstrated in people with aphasia. This finding is consistent with previous reports concerning participants with no cognitive or communication difficulties (e.g., Williams & McCord, 2006) and the strong correlation between the scores from the pen and paper and computerized tests suggests that the computerized format is sensitive to the same

factors as the traditionally administered tests. This result demonstrates that computerized tests could be used to assess people with aphasia, though caution should be exercised as in this study the only scores that were significantly lower than others involved the computer(see figure 3).

There were no significant differences between scores on the two version of the pen and paper conditions which suggests that there was consistency of performance over time and therefore provides some confidence in our interpretation of differences between the pen and paper and computer-based conditions. The lower average scores from the computerized versions of the language tests 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 tests in this study. Although increased difficulty of the computerized tests 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 tests may lead to lower scores is important when considering whether to use them clinically. The difference means that scores from computerized versions of tests should not be compared to scores from the same tests administered in traditional formats, and neither should they be compared to norms collected using the pen and paper version. If pre-therapy baseline scores were collected via computer then the same method should be used when collecting post-therapy measures.

The findings of this study suggest that some people perform better in the presence of a therapist 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: our study showed a different pattern of results across the two tests (see also Schroeders & Wilhelm, 2011). Participants expressed relatively neutral views on this condition: the addition of the researcher in the computer-based task was only significantly better rated on ease of response and whether the participants had enjoyed doing the tasks that way. When comparing the conditions after completing them all, in the method preference questionnaire, the Computer and Therapist condition showed no significant pattern of preference and was rarely selected as the least or most favorite condition. However, participants' qualitative comments indicate 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 alone condition the researcher 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 them for planning or administration tasks. With preparation, the therapist would also be able to use this time even if the client preferred them to remain in the room, as many of our participants did. There may be disadvantages to using computerized tests without the therapist being present however. 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, whilst favored by some, the computer only condition was not popular amongst many of the participants in our study. Whilst comparisons of the experience ratings that participants gave for each condition

were consistently high across the conditions (mean greater than five on the seven-point rating scale), the method preference questionnaire revealed a marked preference for the pen and paper administration of the tests. This mirrors the findings of Ritchie and Newby (1989), who used a similar design and report that college students rated instruction via TV 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 their therapist and the client's engagement with therapy. It is possible that this effect may be mitigated by incorporating – as some therapy programs have done – an avatar whose role is to guide the user through tasks (e.g., Lee & Cherney, 2008). Computer-administered assessment may, however, 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 tests. The failure to find effects in this study might be attributed to the low number of participants, or due to high variability in participants' scores on the tests. However, it should be noted that previous studies which have found an effect of variables such as computer anxiety have tended to be tests of personality and mood (Schulenberg & Yutrzenka, 1999), while 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 be 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. 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, while 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 South East of England, was almost all white and all but four participants were educated past the age of 18. As such, their responses – and views – may not necessarily be representative of all clients with aphasia.

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. Assessment delivered by computers offers well-documented advantages over pen-and-paper assessment. As well as the potential to free up clinician time, it is possible with computer-delivered tasks to make testing more efficient by adapting to the performance of individuals (e.g., discarding items which 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: therapists should critically evaluate computerized tests, should be responsible for ensuring they are appropriately trained 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 on-going form of assessment to monitor progress (Petheram, 2004). This avoids many of the disadvantages of computerized testing by comparing the participant to their 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 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.

Acknowledgements

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

References

Bastiaanse, R., Edwards, S., & Rispens, J. (2002). *VAST: The Verb and Sentence Test.* Bury St Edmunds: Thames Valley Test Company.

Bishop, D. V. M. (2003). *Test for Reception of Grammar.* London: The Psychological Corporation.

Chapelle, C. A. (2008). Utilizing technology in language assessment. In Hornberger, N. H. (Ed.) *Encyclopedia of Language and Education.* 2nd ed. New York, Springer.

Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing, 16*, 49-71.

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*, 295-320.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology, 33*, 593-602.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd edition). New York: Academic Press.

Davis, L., & Copeland, K. (2006). Computer use in treatment of aphasia - a survey of practice patterns and opinions. *Contemporary Issues in Communication Sciences and Disorders, 33*, 138-146.

Dillon, A. (1992). Reading from paper versus screens - A critical review of the empirical literature. *Ergonomics, 35*, 1297-1326.

Gerritsen, M. J. J., Berg, I. J., Deelman, B. G., Visser-Keizer, A. C., & Meyboom-De Jong, B. (2003). Speed of information processing after unilateral stroke. *Journal of Clinical and Experimental Neuropsychology, 25*, 1-13.

Gwaltney, C. J., Shields, A. L., & Shiffman, S. (2008). Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. *Value in Health, 11*, 322-333.

Hill, A. J., Theodoros, D. G., Russell, T. G., Ward, E. C., & Wootton, R. (2009). The effects of aphasia severity on the ability to assess language disorders via telerehabilitation. *Aphasiology, 23*, 627-642.

Ivnik, R. J., Malec, J. F., Tangalos, E. G., & Crook, T. H. (1996). Older persons' reactions to computerized testing versus traditional testing by psychometrists. *Clinical Neuropsychologist, 10*, 149-151.

Kertsez, A. (1982). *The Western Aphasia Battery*. New York: Grune & Stratton.

Kubinger, K. D., Formann, A. K., & Farkas, M. G. (1991). Psychometric shortcomings of Raven Standard Progressive Matrices, in particular for computerized testing. *European Review of Applied Psychology-Revue Europeenne De Psychologie Appliquee, 41*, 295-300.

Lee, J.B., & Cherney, L.R. (2008).The changing "face" of aphasia therapy. *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders, 18*, 15-23.

Mashima, P.A., & Doarn, C.A. (2008). Overview of Telehealth Activities in Speech–Language Pathology. *Telemedicine and e-Health, 14*, 1101-1117.

McDonald, J. L. (2000). Grammaticality judgments in a second language: Influences of age of acquisition and native language. *Applied Psycholinguistics, 21*, 395-423.

McNeil, M.R., & Prescott, T.E. (1978). *Revised Token Test*. Austin, Pro-Ed.

McNeil, M.R., Sung, J.E., Pratt, S.R., Szuminsky, N., Kim, A., Ventura, M., et al., (2008). Concurrent and construct validity of the computerized revised token test (CRTT) and three experimental reading

versions (CRTT-R) in normal elderly individuals and persons with aphasia. Paper presented at the Clinical Aphasiology Conference, Jackson Hole, WY.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests - A metaanalysis. *Psychological Bulletin*, *114*, 449-458.

Merten, T. (1999). Konventionelle und computer-getützte Durchführung von Leistungstets: der Benton-Test (Conventional and computerized test administration: The Benton test). *Zeitschrift für Differentielle und Diagnostische Psychologie*, *20*, 97-115.

Mortley, J., Wade, J., Enderby, P., & Hughes, A. (2004). Effectiveness of computerised rehabilitation for long-term aphasia: a case series study. *British Journal of General Practice*, *54*, 856-857.

Noyes, J., Garland, K., & Robbins, L. (2004). Paper-based versus computer-based assessment: is workload another test mode effect? *British Journal of Educational Technology, 35*, 111-113.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*, 1352-1375.

Palmer, P., Enderby, P., & Hawley, M. (2007).Addressing the needs of speakers with longstanding dysarthria: computerized and traditional therapy compared. *International Journal of Language & Communication Disorders*, *42*, 61–79.

Petheram, B. (2004). Computers and aphasia: A means of delivery and a delivery of means. *Aphasiology, 18,* 187-191.

Ramsberger, G., & Marie, B. (2007). Self-Administered cued naming therapy: A single-participant investigation of acomputer-based therapy program replicated in four cases. *American Journal of Speech-Language Pathology*, *16*, 343–358.

Raven, J., Court, J., & Raven, J. (1995). Coloured Progressive Matrices. Oxford, United Kingdom: Oxford Psychologists Press.

Ritchie, H., & Newby, T.J. (1989). Classroom lecture discussion vs live televised instruction: A comparison of effects on student performance, attitude and interaction. *The American Journal of Distance Education, 2*, 36-43.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational & Psychological Measurement, 71*, 849-869.

Schulenberg, S. E., & Melton, A. M. A. (2008). The Computer Aversion, Attitudes, and Familiarity Index (CAAFI): A validity study. *Computers in Human Behavior, 24*, 2620-2638.

Schulenberg, S. E., & Yutrzenka, B. A. (1999). The equivalence of computerized and paper-and-pencil psychological instruments: Implications for measures of negative affect. *Behavior Research Methods Instruments & Computers, 31*, 315-321.

Schulenberg, S. E., & Yutrzenka, B. A. (2004). Ethical issues in the use of computerized assessment. *Computers in Human Behavior, 20*, 477-490.

Seol, H. (2005). The Critical Period in the Acquisition of L2 Syntax: A Partial Replication of Johnson and Newport (1989). *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 5*, 1-30.

Swinburn, K., Porter, G., & Howard, D. (2004). *Comprehensive Aphasia Test.* Hove: Psychology Press.

Theodorus, D., Hill, A., Russell, T., Ward, E., & Wootton, R. (2008). Assessing acquired language disorders in adults via the Internet. *Telemedicine and e-Health, 14*, 552-559.

Thompson, S. B. N., Ennis, E., Coffin, T., & Farman, S. (2007). Design and evaluation of a computerised version of the Benton visual retention test. *Computers in Human Behavior, 23*, 2383-2393.

Wade, J., Mortley, J., & Enderby, P. (2003). Talk about IT: Views of people with aphasia and their partners on receiving remotely monitored computer-based word finding therapy. *Aphasiology, 17*, 1031-1056.

Wastlund, E., Reinikka, H., Norlander, T., & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior, 21*, 377-394.

Weber, B., Schneider, B., Fritze, E., Gille, B., Hornung, S., Kuhner, T., & Maurer, K. (2003). Acceptance of computerized compared to paper-and-pencil assessment in psychiatric inpatients. *Computers in Human Behavior, 19*, 81-93.

Wertz, R., & Katz, R. (2004). Outcomes of computer-provided treatment for aphasia. *Aphasiology, 18*, 229-244.

Wijndaele, K., Matton, L., Duvigneaud, N., Lefevre, J., Duquet, W., Thomis, M., De Bourdeaudhuij, I., & Philippaerts, R. (2007). Reliability, equivalence and respondent preference of computerized versus paper-and-pencil mental health questionnaires. *Computers in Human Behavior, 23*, 1958-1970.

Williams, J. E., & McCord, D. M. (2006). Equivalence of standard and computerized versions of the Raven Progressive Matrices Test. *Computers in Human Behavior, 22*, 791-800.

Yip, C. K., & Man, D. W. K. (2009). Validation of a computerized cognitive assessment system for persons with stroke: a pilot study. *International Journal of Rehabilitation Research, 32*, 270-278.

Table 1

Characteristics of Participants

Gender	Age	MPO	Ethnicity	Years of education[a]	Previous occupation[b]	RPCM	Computer use (hrs per wk)	Western Aphasia Battery	
								AQ	Aud Comp
M	66	225	W	20	School principal	36	40	65.6	8.5
F	47	44	W	15	Librarian	35	28	87.8	9.2
M	51	32	W	19	Librarian	35	21	54.6	9.4
M	57	65	W	16	Company director	27	15	95	9.2
F	64	36	B	17	Marine biologist	29	14	87.3	8.55
M	41	40	W	16	Graphic designer	33	14	92.7	9.85
F	64	73	W	17	Attorney	28	10	91.6	10
M	53	88	W	16	Futures analyst	36	7	61.6	5.7
M	53	58	W	16	Actor	31	6	84.4	10
F	76	13	W	13	Realtor	32	4	94.3	9.75

M	60	52	W	18	Attorney	27	4	73.5	7.85
F	39	98	B	13	Nurse	33	2	88.6	9.3
M	75	108	W	10	Labor union leader	28	2	71.4	7.7
F	78	218	W	16	Teacher	34	0.75	98	10
F	61	48	W	13	Receptionist	35	0	93.2	9.6

Note: M=male; F=female; MPO=moths post onset; W=white; B=black; RPCM=Ravens Coloured Progressive Matrices; WAB=Western Aphasia Battery; AQ=Aphasia Quotient; Aud Comp=Auditory Comprehension

[a]note that children start school at the age of 5 years in the United Kingdom

[b]none of the participants in the study are currently in employment

Table 2

*Pearson's Correlation Coefficients for the Relationships between One Pen and Paper Condition and the Other Pen and Paper Condition, the Computer Condition and the Computer plus Therapist Condition for Both of the Tasks (*p<.01; **p<.001)*

	Pen and paper	Computer only	Computer and therapist
Sentence comprehension	.89**	.95**	.92**
Grammaticality judgment	.69*	.87**	.78**

Table 3

Average Ratings in the Experience Questionnaire across Three Conditions (1=not at all; 7=very much; asterisks indicate questions where there was a significant difference between the responses for different conditions)

	Pen and paper	Computer only	Computer and therapist
Sound quality?	6.57	6.67	6.53
Picture quality?	6.29	6.67	6.73
Fair questions?	6.04	5.73	5.90
Physically comfortable?	6.86	6.80	6.83
*Easy response?	6.86	5.17	5.87
*Clear instructions?	6.86	6.40	6.47
Fair test?	6.36	6.07	6.53
Nervous?	1.14	1.57	1.43
Confident in results?	5.29	5.83	6.07
*Enjoyed like this?	6.11	5.80	6.50
Happy to repeat?	6.64	6.40	6.70

Figure 1. Screenshot showing some of the instructions for the computerized delivery of the sentence-picture matching task

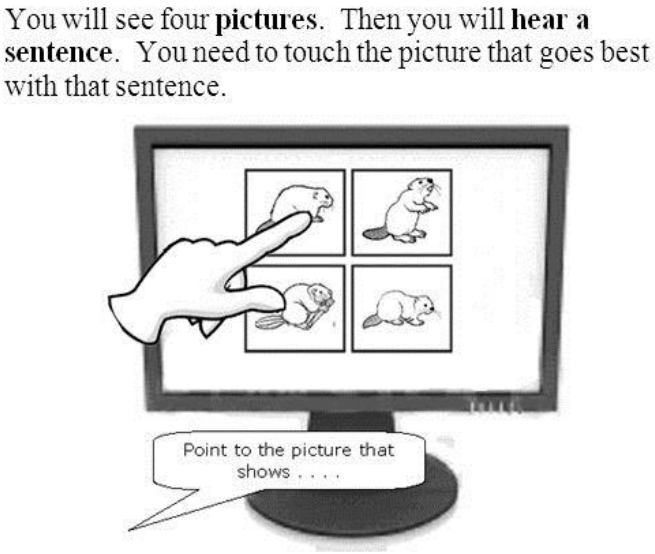


Figure 2. Screenshot showing the computerized delivery of grammaticality judgment task

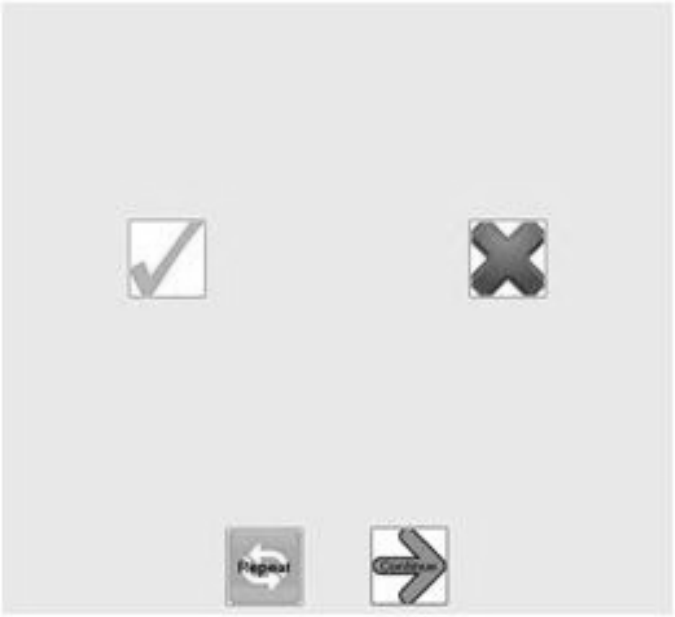


Figure 3. Mean scores for the three conditions for the two tasks (error bars show 95% confidence intervals; asterisks indicate statistically significant differences).

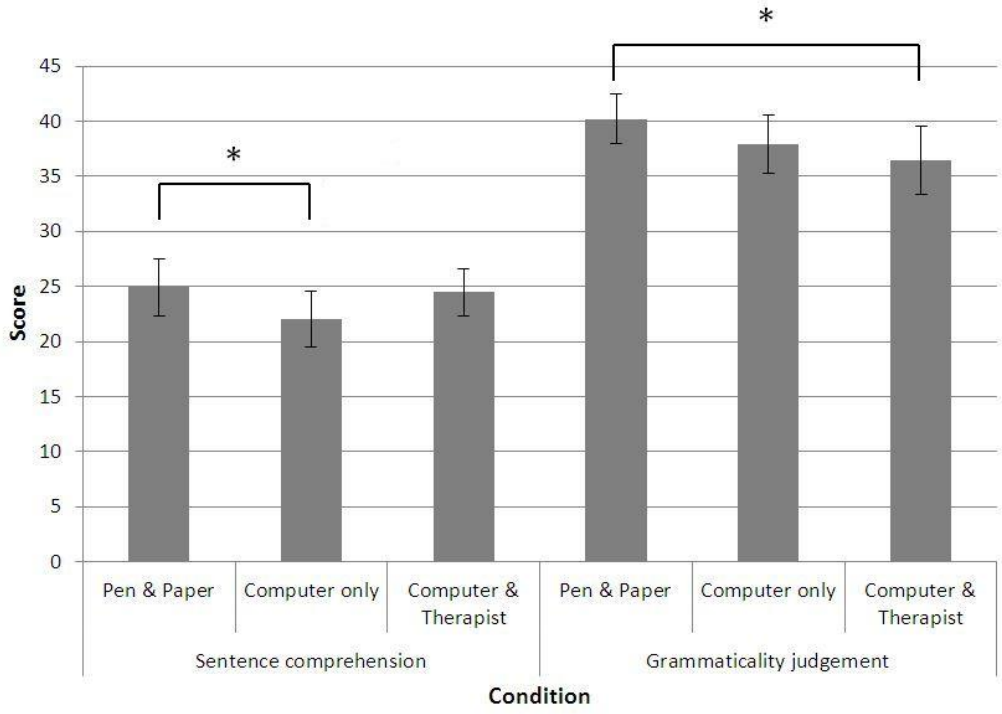


Figure 4. Participants' computer aversion scores from the CAAFI questionnaire

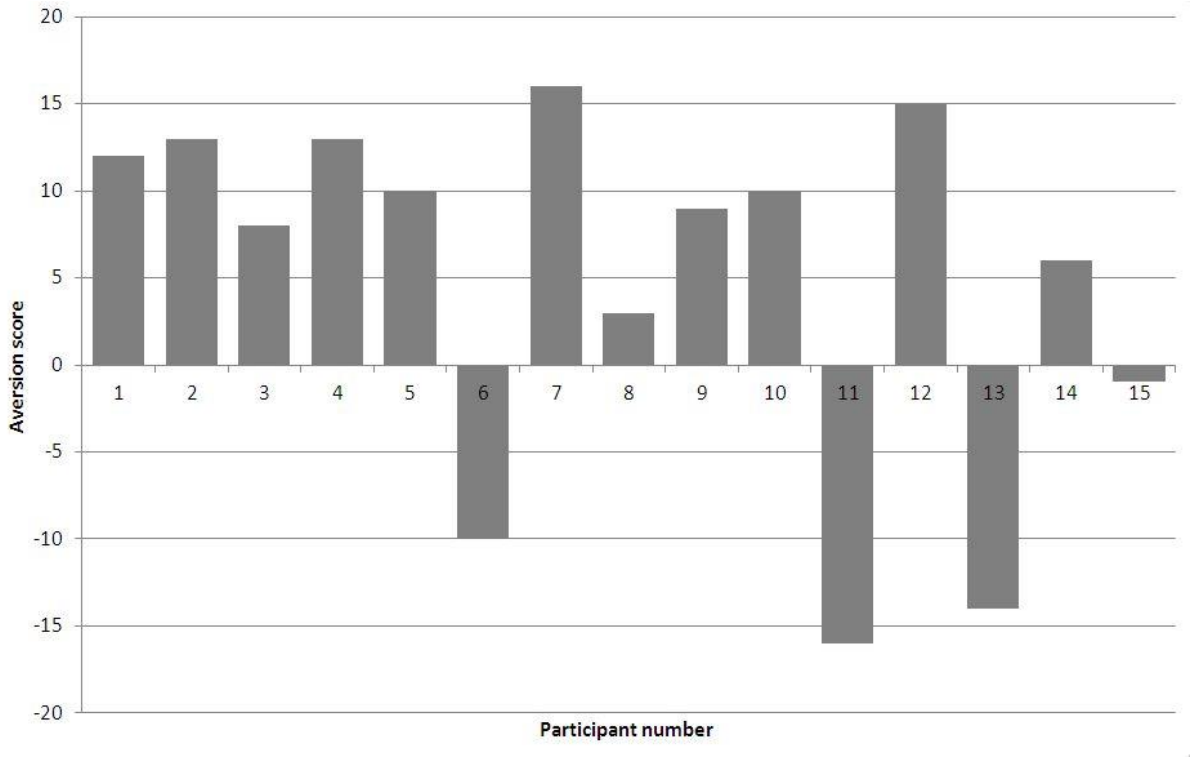


Figure 5. The average time taken to complete each condition, in minutes (error bars are 95% confidence intervals)

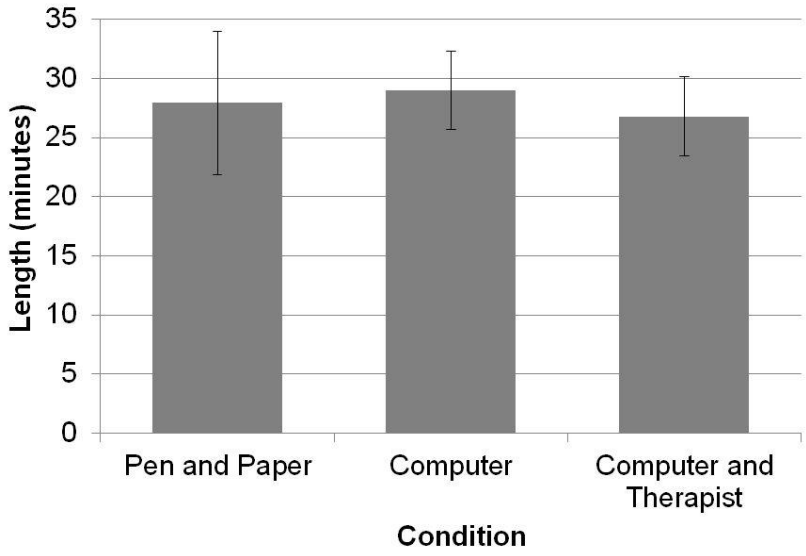
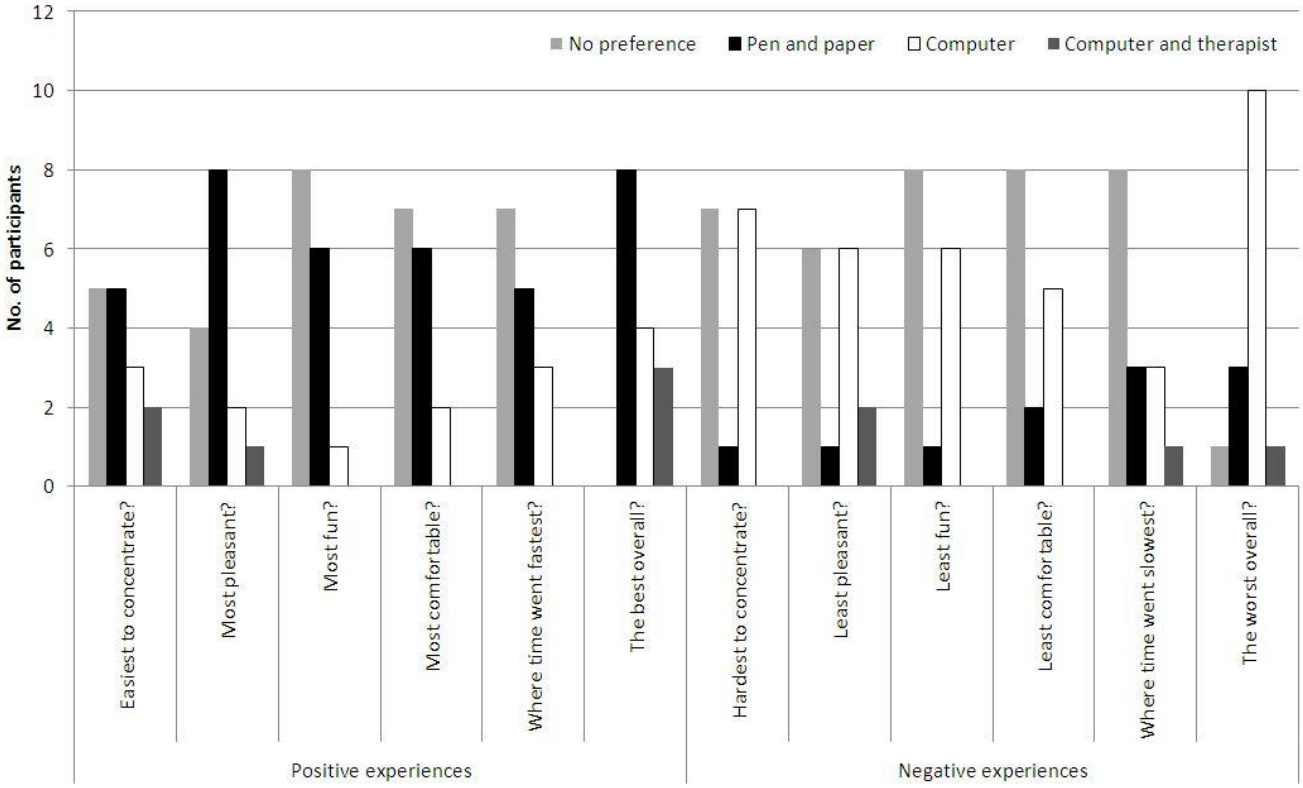


Figure 6. Participants' responses for the method preference questionnaire.



Appendix

Experience questionnaire following each test condition

Think about the test you have just done.

Stimuli

1 Was the sound good quality?

1	2	3	4	5	6	7
Not at all						Very much

2 Were the pictures good quality?

1	2	3	4	5	6	7
Not at all						Very much

3 Were the questions a fair way to test your understanding?

1	2	3	4	5	6	7
Not at all						Very much

Comfort

4 Were you physically comfortable during the test?

1	2	3	4	5	6	7
Not at all						Very much

5 Did you find this an easy way to make responses?

1	2	3	4	5	6	7
Not at all						Very much

Test

6 Were the instructions easy to understand?

1	2	3	4	5	6	7
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?

1	2	3	4	5	6	7
Not at all						Very much

8 Did this way of testing make you nervous? (intimidated)

1	2	3	4	5	6	7
Not at all						Very much

9 Are you confident in the results of this test?

1	2	3	4	5	6	7
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)?

1	2	3	4	5	6	7
Not at all						Very much

11 Would you be happy to do this type of test again?

1	2	3	4	5	6	7
Not at all						Very much