Developing functional outcome measures for unilateral neglect: A pilot study

Audrey Bowen
University of Manchester, Manchester, UK

Elizabeth Gardener, Stephen Cross, Pippa Tyrrell
Salford Royal Hospitals NHS Trust, Salford, UK

Jim Graham
University of Manchester, Manchester, UK

Stroke patients may develop personal neglect, peripersonal neglect or both. Four new measures were tested in a sample of 42 right-handed inpatients (25 male; 17 female, median age 72 years). Participants removed keys from a rack, identified grocery items, washed their face, and cleaned a tray. Prior to this, they were classified as: no neglect (15), personal neglect (8), peripersonal neglect (7), and both personal and peripersonal neglect (12). The sensitivity and specificity of each new measure was determined by agreement with the classification. Test–retest reliability was determined using weighted kappa statistics or limits of agreement. Four occupational therapists (OTs) rated videos of the face and tray measures, and software was developed to measure objectively time spent and area covered on Face and Tray.

Keys and Grocery had high specificity, good reliability but poor sensitivity. For the OTs’ video ratings, there was good and moderate inter-rater reliability on Tray and Face respectively for area covered, but not time spent. Intra-rater

Correspondence should be addressed to Audrey Bowen, Human Communication and Deafness, Faculty of Education, University of Manchester, Oxford Road, Manchester, M13 9PL, UK. Tel: +44 (0) 161 275 3401, Fax: +44 (0) 161 275 3373, Email: audrey.bowen@man.ac.uk.

We would like to thank the patients and clinicians of the North West Stroke Services for their support with this project. We are especially grateful to the four occupational therapists who participated in the video rating part of the study, and to Kola Babalola who helped develop the software. This study was funded by the NHS Executive North West as part of the Research and Development Fund programme. The project grant-holders were: A. Bowen, E. Gardener, P. Tyrrell, S. Cross, R. Collis, and R. Tallis. Beverley Jones was the research assistant.
reliability was also better for area than time on Tray. However, the validity of Face and Tray themselves is currently inadequate. A longitudinal study is proposed to modify the measures, increase their sensitivity and evaluate their ability to monitor change over time.

**INTRODUCTION**

Unilateral spatial neglect is a commonly occurring cognitive deficit following stroke, although the incidence and prevalence rates vary widely (Bailey & Riddoch, 1999; Bowen, McKenna, & Tallis, 1999). Neglect can reduce a person’s ability to look, listen or make movements towards one side of that person’s environment, and disrupt activities of daily living (ADLs), increasing dependence on others. Considerable nursing and therapy resources are directed towards neglect, yet there is no evidence for the lasting effectiveness of any intervention on ADLs (Bowen, Lincoln, & Dewey, 2003). Recent advances in both the understanding of neglect and the development of rehabilitation techniques (Robertson et al., 2002; Schindler et al., 2002) mean that functionally meaningful and psychometrically sound outcome measures are now needed for randomised controlled trials of therapy.

Existing measures may lack relevance to ADLs or specificity to neglect-related behaviour. For example, bisection and cancellation tasks are most likely to be measuring at the level of “impairment”, as defined by the World Health Organisation (WHO, 2001). Impairment level measures may play a useful role in screening for the presence of neglect (i.e., diagnostic measures) but changes at a more ecologically meaningful level are required for evaluating the effectiveness of interventions. By this we mean that outcome measures should be at the level of “activity limitations” or “participation restrictions”. In addition, measures must have adequate specificity. Some assessments may provide too imprecise a measurement to show a change in functioning following the rehabilitation of neglect. The Barthel Index (Shah, Vanclay, & Cooper, 1989), for example, is used to reflect changes on gross activities such as feeding and mobility.

Several assessment instruments currently exist, which were developed specifically for neglect and include measures based on everyday functional activities. For example, the Behavioural Inattention Test (BIT) battery contains several subtests such as, reading a menu, telling the time, and identifying objects for personal hygiene (Wilson, Cockburn, & Halligan, 1987). Although the BIT is sometimes used as an outcome measure it was intended for diagnosis and to inform the rehabilitation team’s programme for that individual. While useful for diagnosis and informing rehabilitation, there is as yet no evidence that it is responsive to the possible changes following rehabilitation for neglect. Once again this returns to the important distinction between
measures suitable for diagnosis and those for outcome. Furthermore, used as a whole, the BIT contains some impairment level (e.g., star cancellation) tests that may not be appropriate as outcome measures. The entire BIT is also time-consuming to administer.

There are three categories of other possible neglect-specific functional outcome measures. First, a 19-item Subjective Neglect Questionnaire (SNQ) has been used to elicit families’ (and patients’) subjective accounts of the impact of neglect on everyday activities, such as misreading words, finding the wheelchair veering off, and bumping into furniture (Towle & Lincoln, 1991). The questionnaire emphasises peripersonal neglect and there is only one item (to do with dressing) that may measure personal neglect. Another possible limitation is that many of the items were also reported as causing difficulty for patients without neglect. In fact more “non-neglect” than “neglect” relatives reported that the patient bumped into furniture (27% and 19%, respectively). The authors themselves recommend that further work is required to investigate the SNQ’s usefulness in planning and evaluating treatment.

A second category is the approach developed by the French collaborative neglect study group (Azouvi et al., 2002). They favour a 10-item checklist, the Catherine Bergego Scale (CBS), completed by a specialist occupational therapist (OT) who knows the patient well and has observed his or her performance in real-life situations (Azouvi et al., 2003). The CBS contains items similar to those in the SNQ (e.g., collides with people or objects, such as doors or furniture, on the left side). It does however provide a more balanced description of neglect in that it contains items likely to be affected by personal, peripersonal (and possibly extra-personal) neglect. The French group argues that observations of real-life situations may pick up on neglect in a way that is missed in “simulated tasks”, when the patient knows he or she is being assessed. They base this on the argument that neglect occurs during automatic, but not necessarily voluntary, actions due to a relative sparing of voluntary orientation of attention (Seron, Deloche, & Coyette, 1989).

The CBS appears promising, at least in situations where a specialist OT is available. Its value in research evaluations of interventions is not yet known. Further work must be done on its responsiveness to change over time. Additionally, there is a need to demonstrate the feasibility of ensuring that the OT who knows the patient well enough to comment on each of the CBS’s items is somehow blinded to randomised allocation within a trial. One further important point concerns the CBS’s grouping together of items that measure different types of neglect. The most recent study suggests these items are a “homogeneous construct” (Azouvi et al., 2003). However, double dissociations between neglect of personal and of peripersonal space clearly exist (Beschin & Robertson, 1997). These, and other dissociations, support the widespread belief that neglect is not a “unitary phenomenon” (Bailey, Riddoch, & Crome, 2000).
Given this heterogeneity, the rehabilitation of neglect is most likely to be effective (and effectively demonstrated) if the intervention strategy and outcome measure are tailored to the specific disorder. For example, a personal neglect outcome measure is best suited to evaluate the effectiveness of a personal neglect intervention. We would argue that, for evaluation purposes, there is a need for type-specific functional outcome measures. For the purposes of collecting research evidence, these need to be valid and reliable when completed by researchers who do not know the patient. Once the evidence exists and interventions are implemented into standard practice then measures such as the CBS and SNQ may be useful in monitoring progress. Incidentally, in view of the automatic versus voluntary distinction discussed above, these latter measures may also be better suited to determining incidence and prevalence rates.

The third category of outcome measures consists of the simulation of everyday tasks. Personal grooming tasks have received the greatest amount of research attention since they were first advocated by Italian researchers (Pizzamiglio, Judica, Razzano, & Zoccolotti, 1989; Zoccolotti & Judica, 1991). Methods of scoring the performance of simulated face-shaving or applying face powder, and hair-combing have since been greatly improved (Beschin & Robertson, 1997; McIntosh, Brodie, Beschin, & Robertson, 2000). However, as with the Swedish approach of scoring performance on a simulated cake-baking task (Tham & Tegner, 1996) these are somewhat limited by being sex specific (shaving, baking). Hair-combing seems to be based on the assumption of a symmetrical hairstyle. The latest scoring method (percent bias index) yields a positive score for a rightward bias, which is thought to indicate left neglect. However, there may be other, more practical reasons for extra combing on one side, such as a side parting.

Aims and objectives

The aim of this study was to develop psychometrically sound outcome measures. Four novel outcome measures were developed from functional activities that were unimanual, familiar, unisex, quick and easy to complete and did not require the participant to mobilise. Separate measures were developed for personal neglect (Face Washing) and peripersonal neglect (Tray Wiping, Grocery Naming, Key Removing). Each measure is described below.

The specific objectives were:

1. To determine the validity of the four new measures by examining agreement with a lengthier existing clinical diagnostic battery.

2. To determine the test–retest reliability of each new outcome by comparing the results of two separate measures taken from participants on the same day.

3. To develop computer software to measure objectively the percentage of area covered and time spent by the participant’s hand on the left side on the
Face/Tray measures, and to ensure these computerised ratings themselves had adequate intra-rater reliability for both area and time.

4. To determine the intra-rater and inter-rater reliability of OTs’ ratings of Face/Tray measures.

**METHOD**

**Participants**

An opportunity sample was recruited from people with stroke who were inpatients at seven NHS Trust Hospitals in the North West of England according to the following criteria. Participants: had a recent right hemisphere stroke, verified by clinical judgement of the consultant and computerised tomography scan (when available); had sufficient sitting balance to sit independently or supported in a chair or in bed; were right handed, and had sufficient comprehension to give informed consent. Excluded were those with evidence of persisting neglect from an old stroke where damage occurred in the opposite cerebral hemisphere to the new stroke; inadequate sitting balance; and those who were left hand dominant before onset of stroke; had a nasogastric tube in situ, or insufficient English comprehension to give voluntary informed consent.

The sample size was determined by the numbers admitted over the short period of time available for this pilot study. Participants thought to be potentially suitable for the study, were identified by health professionals. The participation rate was high. Eight patients refused to be screened for eligibility. Forty two were eligible and consented to participate. As shown in the final column of Table 1, the majority were male (60%), their median age was 73 (the youngest/oldest person was 25/98), they were seen about six weeks post-onset at which time they still had a fairly poor median Barthel score (43).

**Procedure**

All 42 were administered an existing clinical diagnostic battery of neglect assessments, an interview and observations to allocate them to one of four possible diagnostic groups: no neglect, personal neglect only, peripersonal neglect only, both personal and peripersonal neglect, to validate the novel outcome measures (Objective 1).

Four *diagnostic* assessments were chosen. These had demonstrated their utility in detecting the presence of personal or peripersonal neglect in previous studies and for their ease of use with older people with post-stroke disabilities. Two were to detect personal neglect. The first was a modified version of the Fluff Test (Cocchini, Beschin, & Jehkonen, 2001). In our modified version 12 small fluffy balls with a sticky backing were attached to the participant’s
clothing on the upper half of the body only. This eliminated the need for participants to get out of bed or to lean forwards to their lower body. Participants scored one point for each fluff ball removed.

The second personal neglect assessment was the Bisiach test (Bisiach, Perani, Vallar, & Berti, 1986). The participant was asked to sit, with both hands resting on a table and their eyes closed. The researcher touched the participant’s right hand and said “with this hand, touch your other hand”. This was scored from 0 to 3 as follows:

0 (no neglect) the patient promptly reached for the target.
1 (mild neglect) the target was reached with hesitation and search.
2 (moderate neglect) the search was interrupted before the target was reached.
3 (severe neglect) no movement towards the target was performed.

The two peripersonal assessments were Star Cancellation from the Behavioural Inattention Test (Wilson et al., 1987) and a modified version of the Baking Tray Test (Tham & Tegner, 1996). The Star Cancellation Test involved asking the participant to use a pen to cancel out each of the 54 small stars printed on a sheet of A4 paper and to avoid cancelling distractor stimuli. The number omitted on the left and right is calculated. Typically people with peripersonal left neglect will cancel fewer on the left side.

The fourth assessment was a modified version (Wenman et al., 2003) of The Baking Tray Test. The participant was seated at a table. The centre of the long edge of a real baking tray (containing 16 salt-dough buns in paper cases, evenly spread across its surface) was placed directly in front of participants’ mid sagittal plane. Using their right hand, participants removed as many of the buns as they could and placed them in a plastic container on their lap. No prompts were given. The number removed from the left and right side was recorded.

The order of the four assessments was counterbalanced across participants. Interpretation of assessment scores took into account behavioural observations during testing. These were discussed with the principal investigator and agreement was reached on allocation to the diagnostic groups. Fifteen (36%) participants showed no indication of neglect. The majority (12) of those with neglect had both an impairment of their body space (personal neglect) and the space within grasp (peripersonal neglect). Eight had only personal neglect and seven had only peripersonal neglect. The three neglect groups are merged in Table 1.

Novel outcome measures. Each participant was seated at a table with the exception of two who needed to be assessed sitting in bed. Materials were presented centrally to the participant. Standardised instructions were given for each
outcome measure. The measures were completed in a counterbalanced order between participants. The complete set of measures was repeated after a short break (and always within 24 hours) to determine test–retest reliability (Objective 2).

1. Key Removing—The participant removed keys from a keyrack (915 mm in length) containing nine keys hanging on hooks 110 mm apart. There was no time limit. The researcher noted the starting point and the position of any neglected keys.

2. Grocery Naming—The participant named common grocery items in a kitchen cupboard. There was no time limit. The cupboard contained 14 grocery items, displayed horizontally on two shelves with seven items per shelf. For ease of transport a cupboard frame was used (49.5 cm × 44.5 cm) containing a lifesized colour photograph of the contents. The researcher noted the starting point and the position of any neglected grocery items.

3. Face Washing—The participant wiped his or her face for 20 seconds using a lightly dampened small blue sponge held in the right hand. This was videoed and any bias in the area covered or proportion of time spent on the left was later scored (calculated by the software and rated by the OT, see below).

4. Tray Wiping—The participant wiped a standard size tray (40.5 cm × 34 cm) for 20 seconds using a dampened white nylon kitchen sponge (approx. 10 cm × 5 cm). This was videoed and any bias in the area covered or proportion of time spent on the left was later scored (calculated by the software and rated by the OT, see below).

### TABLE 1
Demographic and clinical descriptives at recruitment

<table>
<thead>
<tr>
<th></th>
<th>No neglect n = 15</th>
<th>Any type of neglect n = 27</th>
<th>All participants n = 42</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex: male</td>
<td>9 (60%)</td>
<td>16 (59%)</td>
<td>25 (60%)</td>
</tr>
<tr>
<td>Median age in years (IQR)</td>
<td>74 (70,79)</td>
<td>71 (64,80)</td>
<td>73 (66,80)</td>
</tr>
<tr>
<td>Median Barthel Index (IQR)</td>
<td>55 (35,72)</td>
<td>30 (19,60)</td>
<td>43 (24,60)</td>
</tr>
<tr>
<td>Type of stroke</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TACS</td>
<td>4 (27%)</td>
<td>19 (70%)</td>
<td>23 (55%)</td>
</tr>
<tr>
<td>PACS</td>
<td>9 (60%)</td>
<td>5 (19%)</td>
<td>14 (33%)</td>
</tr>
<tr>
<td>LACS</td>
<td>2 (13%)</td>
<td>2 (7%)</td>
<td>4 (10%)</td>
</tr>
<tr>
<td>SAH</td>
<td>0</td>
<td>1 (4%)</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Median number of days since onset of stroke (IQR)</td>
<td>61 (20,78)</td>
<td>38 (27,80)</td>
<td>42 (24,79)</td>
</tr>
</tbody>
</table>

SAH = subarachnoid haemorrhage, T/P ACS = total/partial anterior circulation syndrome, LACS = lacunar syndrome.
The Face/Tray videos were used to develop customised software to provide an objective measure of the percentage of the area covered, and time spent, by the participant’s hand on the left side. The videos were digitised and run through the programmes. These pilot versions of the face and tray software required some subjective decisions by the researcher (e.g., manually specifying the corners of the tray). Therefore, they were run twice to examine intrarater reliability of the use of the software (Objective 3).

**Occupational therapists’ ratings.** The (non-digitised) videos of all 42 participants performing Tray Wiping and Face Washing were rated by four OTs with experience of stroke rehabilitation. This was done under controlled conditions at the university. Two of the OTs worked in hospitals and two in the community. They were 3–9 years post-qualification and had between 1 and 6 years experience in stroke rehabilitation. All felt sufficiently experienced to assess for neglect.

Each OT independently rated (see below) all 42 participants’ first performance on the tray (Objective 4), followed by all participants’ second tray performance (Objective 2). In order to examine OTs’ intrarater reliability this was then repeated. The OTs returned to the university for a second afternoon where they rated all the Face videos.

The ratings were of (1) area covered and (2) proportion of time spent on the left side of the face. After watching a 20 second video clip, OTs used a felt pen to shade in the areas washed/wiped on an outline of a face/tray on a sheet of A4 paper. The researcher later scored the OTs’ rating, overlaying a grid printed transparency and noting which squares were shaded or neglected. To determine OTs’ ratings of any bias in the proportion of time spent by the participant on the left or right of the Face/Tray, the OTs ticked one of three boxes. The three options were “more time spent on the left”, “more time spent on the right” or, “neither—both the same”.

**PLAN OF STATISTICAL ANALYSES**

To demonstrate whether the new measures (Key Removal and Grocery Naming) agreed with the clinical assessment (validity) we considered the sensitivity and specificity at different cut-offs (Objective 1). Sensitivity is the proportion of patients with peripersonal neglect correctly identified. Specificity is the proportion of patients without peripersonal neglect that are correctly identified.

Test–retest reliability of Key Removal and Grocery Naming was assessed using a weighted kappa (Objective 2). The kappa statistic measures agreement beyond chance and can take values between −1 and +1, 0 indicating agreement no better than chance. A kappa value over 0.75 signifies very good
agreement. The weighted kappa weights the disagreements depending on the size of the discrepancy. The weight used was the absolute error weight. Where possible, 95% confidence intervals for kappa statistics were calculated. Kappas without a confidence interval should be treated as an estimate.

The Face/Tray software’s intra-rater reliability was determined. The results of the same video being run through the software twice by the researcher was examined using limits of agreement (Bland & Altman, 1986) examined (Objective 3). Then, the mean of both computerised ratings was used to determine the validity of Face/Tray. Dot plots were used to see if the distribution of the percentage of time spent on the left (or the percentage of area covered on the left) differed for patients with and without the relevant type of neglect as determined on the existing diagnostic battery. Logistic regression models were fitted, to establish whether there was a significant relationship between neglect diagnosis and the outcome measure (Objective 1). Limits of agreement were then used to assess test–retest agreement for Face/Tray (Objective 2) when objectively measured with the software.

The occupational therapists rated “per cent of total area that was on the left” for the Tray Wiping and the Face Washing outcome measures. Because many ratings were exactly 50% it was not reasonable to model “per cent of total area that was on the left” as continuous data. Therefore, “per cent of total area that was on the left” was categorised: “More of the area covered was on the left”, “More of the area covered was on the right” and “Same amount on both sides”. Time spent was categorically rated as “more on left”, “more on right”, or “neither—both the same”. Inter-rater agreement was assessed by calculating a combined Kappa statistic from the first rating of the four OTs for the first time the patient did the task (Objective 4). This combined Kappa is the weighted average of the individual Kappas, which are calculated separately for each rating (e.g., more on left) against the other two ratings (more on right, same on both). Intra-rater agreement was assessed by calculating Kappa statistics for each of the therapists. Only the ratings for the first time the patient did the activity were used.

RESULTS

Key Removal

Choosing nine keys as the cut-off maximises sensitivity, but it is still low (47%). More than half of the patients clinically diagnosed as having peripersonal neglect removed all nine keys, so would be incorrectly diagnosed as having no neglect. The specificity with nine as the cut-off is considerably higher (96%) than the sensitivity. One patient who was not clinically diagnosed as having neglect removed less than nine keys, so would be incorrectly diagnosed as having neglect.
The weighted Kappa statistic was 0.64 (95% CI 0.42 to 0.87) indicating good agreement (test–retest reliability). The overall proportion of agreement was 79%, most agreement occurred when all nine keys were removed. Three patients removed fewer keys the second time than the first (all three saw one less). Six patients removed more keys the second time than the first (two patients one more, two patients two more, one patient three more, and one patient four more).

Grocery Naming

Using 14 as the cut-off gives the highest sensitivity (53%) and specificity (100%). Although the test is specific (all patients that were diagnosed clinically as not having peripersonal neglect saw all 14 items) the sensitivity is low (of the patients that were clinically diagnosed as having peripersonal neglect nearly half of them saw all 14 items, so would be incorrectly classified as not having neglect).

The weighted kappa statistic of 0.80 (95% CI 0.55 to 1) indicated very good agreement. The overall proportion of agreement was 81%, most (71%) agreement occurs when all 14 items were seen. Five people saw fewer items the second time than the first (three people saw one less and two people two less). Three people saw more items the second time (two people saw two more and one person saw one more).

Face Washing

Figure 1 shows the actual differences between the two occasions the software was used to rate the participants’ first performance of Face Washing when area covered was measured. The limits of agreement are acceptable (for both area and time) suggesting good intra-rater reliability for the face software (see Table 2). However, the accuracy is uncertain as there is little spread (e.g., Figure 1). Resource difficulties in this pilot funded project meant that only half the 42 Face data sets were randomly selected for digitisation and analysis.

From logistic regression analysis, the percentage of area covered on the left side of the face is not related to personal neglect as determined by the existing diagnostic battery, \( p = .587 \). As Figure 1 shows, those in the personal neglect group (circles) spent around 50% of their time on the left side of their face. There was greater variation in the no neglect group (crosses). A similar pattern was seen for time spent, and the percentage of time spent is not related to personal neglect either, \( p = .644 \).

Figure 2 shows the actual differences between the two occasions the patients performed Face Washing, when time spent was measured (test–retest reliability). The limits of agreement are wide (−27.04 to 26.54) suggesting little reliability. There was a very similar finding for area covered (−16.71 to 17.55).
Tray Wiping

Resources were available to digitise all 42 participants’ Tray videos. The limits of agreement suggested reasonable intra-rater reliability for the Tray software’s measurement of area covered and time spent (see Table 2). (Time spent has wider limits of agreement, largely due to data from one participant which showed an extreme difference.) However, the validity of the Tray task is poor. From logistic regression analysis, neither the percentage of time spent

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Mean difference (software 1–software 2)</th>
<th>Limits of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per cent area covered on left side of Face</td>
<td>-0.38</td>
<td>-4.18 to 3.41</td>
</tr>
<tr>
<td>out of total covered area</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per cent time spent washing left of Face</td>
<td>0.54</td>
<td>-3.88 to 4.96</td>
</tr>
<tr>
<td>out of total time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per cent area covered on left side of Tray</td>
<td>-0.8</td>
<td>-3.7 to 2.1</td>
</tr>
<tr>
<td>out of total covered area</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per cent time spent wiping left of Tray</td>
<td>-0.6</td>
<td>-6.9 to 5.6</td>
</tr>
<tr>
<td>out of total time</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Resource difficulties meant that half the 42 Face data sets were randomly selected for digitisation and analysis.

Figure 1. Percentage of area covered on left of Face out of total area, (intra-rater reliability) n = 21.
(p = .510) wiping the left side of the Tray nor the area covered (p = .429) is related to peripersonal neglect as determined on an existing diagnostic battery. Furthermore, the Tray’s test–retest reliability is poor with wide limits of agreement for time spent (−17.83 to 15.00) and area covered (−10.14 to 8.66).

**OT ratings of Face and Tray**

**Face.** For time spent on the Face, there was inter-rater agreement between all four OTs’ ratings for 22 (52%) of the participants, 1 of these was “more on the left”, 19 were “both the same” and 2 were “more on the right”. The combined Kappa was fairly poor (0.39). Separate Kappas were 0.42, 0.37, and 0.41 for “more on left”, “both the same” and “more on right”, respectively. The Kappa statistics (95% CI) for the OTs’ intra-rater agreement of time spent ranged from 0.34 (0.10, 0.58) to 0.62 (0.36, 0.89).

For area covered on the Face, there was inter-rater agreement between all four OTs’ ratings for 27 (64%) of the patients, 3 of these were “more on the left”, 19 were “both the same” and 5 were “more on the right”. The combined Kappa was good (0.65). Separate Kappas were 0.54, 0.77, and 0.57 for “more on left”, “both the same” and “more on right”, respectively. The Kappa statistics (95% CI) for the OTs’ intra-rater agreement of area covered ranged from 0.37 (0.17, 0.56) to 0.65 (0.42, 0.87).

**Tray.** For time spent on the Tray, there was inter-rater agreement between all four OTs’ ratings on 23 (55%) of the participants, 21 were agreed to be “both
the same” and 2 “more on the right”. The combined Kappa was poor (0.30). Separate Kappas were 0.08, 0.34, and 0.33 for “more on left”, “both the same” and “more on right”, respectively, indicating that there was much less agreement for “more on left” than the other ratings. The Kappa statistics (95% CI) for the OTs’ intra-rater agreement of time spent ranged from 0.19 (−0.06, 0.45) to 0.43 (0.21, 0.66).

For area covered on the Tray, there was inter-rater agreement between all four OTs’ ratings for 38 (90%) of the patients, 33 of these were “both the same” and 5 “more on the right”. The Kappa was very good (0.80). Note separate Kappas were not needed because no rating was “more on the left”. The Kappa statistics (95% CI) for the OTs’ intra-rater agreement of area covered ranged from 0.53 (0.22, 0.83) to 0.81 (0.50, 1).

DISCUSSION

This study found that the new measures showed promise but, as they stand, they have insufficient sensitivity for use as outcome measures. Before summarising the results and proposals to improve sensitivity, we should first discuss the important issue of “gold standards” in measurement. This discussion does not change the fact that the measures lacked sensitivity but it does provide an important context in which to consider results related to the validity of the new measures.

Ideally, in a study of this type, one would have gold standard outcome measures against which to test the validity of the new outcome measures. As described in the introduction, this was not the case in the present study as gold standard outcome measures do not yet exist for neglect. In fact, if they did, it would be difficult to justify developing new measures. Our comparator of validity was a classification (no neglect, personal neglect only, peripersonal only, both) derived from a battery of existing diagnostic assessments and observation of qualitative aspects of the person’s behaviour during assessment.

In some cases the classification was difficult to do, in that some diagnostic assessments in the battery contradicted each other. However, we feel these classifications add value, as long as we make clear that certain “failures” of the new outcome measures are their performance relative to the adequacy of the diagnostic measures. As described in the introduction, there is an important distinction between diagnostic and outcome measures. Each has a separate purpose. It could be argued that measures designed as diagnostic instruments are not adequate for comparison with potential outcome measures as the latter must go beyond diagnosis. However, we would argue that outcome measures must at least be as precise as diagnostic assessments. In fact a long-term goal of developing new outcome measures might be that they will ultimately serve a joint purpose: diagnosing (and informing the choice of therapy), before going on to monitor responsiveness to change (post-therapy outcome).
This pilot study has not sought to examine responsiveness to change but whether a longer-term study was justified. We believe we have shown that it is.

To summarise the results, analyses of the new outcome measures (and software for research purposes) showed promise. Some measures (e.g., Keys and Grocery) had high specificity and good reliability. However, further development is required to increase their poor sensitivity to an acceptable level. These tasks could easily be modified. Increasing the task difficulty (and therefore the attentional demands) may elicit neglect behaviour. This prediction is based on the established evidence that the degree of neglect is highly positively correlated with difficulty on tests of attention such as Digit Span and the Paced Auditory Serial Addition Test (PASAT) (Robertson, 1993). There is a growing body of evidence to suggest that many people with left neglect also have a non-lateralised attentional deficit (e.g., of sustained attention), especially those with persisting neglect (Maguire & Ogden, 2002; Manly, 2002). Rehabilitation approaches aimed at sustained attention training are beginning to emerge (Wilson & Manly, 2003). The modifications proposed are to increase the number of groceries and shelves in the cupboard to require greater visual scanning and sustained attention. Less common grocery items could be used to increase the level of cognitive processing (object recognition, naming).

Development of customised software for the other two measures (Face and Tray) was successful, but the pilot versions are time-consuming to use and could be improved. These provided a reliable measure of time spent and area covered on the left of the Face and Tray. Area covered appears more promising and will be developed instead of time spent in a subsequent study. This study would also improve the software in terms of speed of data processing and accuracy of the output.

However, before improving the software, the inadequate validity of the Face and Tray measures themselves must be addressed. In their current form these measures also lacked precision. The present study’s findings suggested two modifications to improve validity. Altering the instructions so that participants wash or wipe for a shorter, more realistic period of time may elicit the neglect behaviour. Our concern to capture a full 20 seconds per participant may have been unnecessary and may have increased the artificiality of the task. Many people reported that 20 seconds was far longer than they would usually take to wash their face. The researcher’s encouragement to persist may have cued them to attend more carefully than they would normally and therefore obscured their disability.

A future study would capture naturalistic face washing. Videoing is essential but our patients’ awareness of looking straight into the camera and lighting may have raised arousal, improved alertness and temporarily masked their personal neglect. Future videoing would make use of the miniature webcams that are now available and avoid instructions that focus the patient’s attention on the task. It may also be useful to measure washing of the left arm.
(using the right, non-paretic hand to do so) since this requires a considerable movement beyond the body’s midline.

We received valuable feedback from the OTs in the study. They were positive about the value of the Face Washing task, although they found it difficult to rate for 20 seconds per person. They suggested other methods of rating the videos that a future study could explore. Alternatively, the software itself could be developed to produce a quick and objective quantification of the amount of neglect present on the video. Regarding the Tray Washing task, the OTs suggested changing the tray to one without a rim. They felt that the rim cued horizontal visuomotor attention in people who would otherwise have shown neglect (as they traced their hand along the rim). A future study may consider there is no need for a tray task if (as discussed later) a modified grocery (removal) task is used instead as a visuomotor peripersonal task.

For the OTs’ video ratings, we found good and moderate inter-rater reliability on the Tray and Face tasks, respectively, but only for area covered (i.e., not for time spent). Intra-rater reliability was also better for area than time on the Tray. Further analyses of the OT ratings were not carried out. The poor validity of the Tray and Face tasks when compared to the diagnostic battery meant that it was not sensible to analyse the validity of OT ratings. A future study would need to examine these types of validity once the modified tasks themselves had been shown to have adequate sensitivity, if an OT rather than software rated measure is required. We originally intended to develop OT/nurse-rated measures to reduce costs. However, most stroke units would have access to a video recorder and personal computer. It may be worth trying to refine the software to make it quick, automatic and capable of running on a personal computer with limited processing speed.

In addition to the above proposals to increase the sensitivity of Grocery Naming (e.g. to test the hypothesis that increasing the tasks’ attentional demands may elicit neglect) a subsequent study could also aim to modify this task so that it (1) replaces Key Removal and (2) functions as a measure of both visual and visuomotor neglect. The replacement, Grocery Removal, would be a modified version of the existing (non-motor) Grocery Naming task and would involve reaching and fine finger control (i.e., grasping and lifting of grocery items). By using the same set of stimuli (grocery items) in two different tasks (Grocery Naming and Grocery Removal) we would provide measures tailored to different types of neglect, visual and visuomotor respectively.

With hindsight, the diagnostic test “Menu-reading” from the BIT would be a better comparator for the visual neglect part of the grocery task. Menu-reading (unlike Star Cancellation and Baking Tray used in the present study) only requires a visual response. In fact the reported poor sensitivity of the “visual” Cupboard in the pilot study may be due to the fact that it was unfairly compared to visuomotor tasks.
To conclude, this pilot study supports the feasibility and value of carrying out a full-scale study on modified versions of these outcome measures. There is as yet no evidence with which to recommend changes to practice (i.e., the measures currently used to choose and evaluate therapy approaches). Given the conclusions of the recently published (Bowen et al., 2003) Cochrane Review of neglect (that the existing evidence base for the rehabilitation of neglect is inadequate) we believe further work on refining new outcome measures would be of value to users, providers and commissioners of clinical services, and to researchers interested in evaluating new and existing interventions. This would complement those reviewed in this paper and in some cases provide a measure where no suitable measure currently exists.

REFERENCES


Manuscript received August 2003 Revised manuscript received October 2003