# The Test-retest and Inter-rater Reliability of the Structured Observational Test of Function (SOTOF)

This is an edited version of chapter 7, from Laver's PhD thesis for the SOTOF 2nd edition test manual (Laver-Fawcett, 2016)

**Reference for original chapter:**
Laver, A. J. (1994). The Test-retest and Inter-rater Reliability of the Structured Observational Test of Function (SOTOF) (Chp. 7). In: *The development of the Structured Observational Test of Function (SOTOF).* PhD thesis. Guildford: Department of Psychology, University of Surrey. (Available from the College of Occupational Therapists library in London, UK)

**Reference for this version:**
Laver-Fawcett, A. J. (2016) The Test-retest and Inter-rater Reliability of the Structured Observational Test of Function (SOTOF). York: York St John University.

## Summary

The purpose of this study was to evaluate the test-retest and inter-rater reliability of the original version of the SOTOF. The method involved the examination of the correlation between (1) scores obtained by pairs of occupational therapist raters scoring the same administration of the SOTOF to an older person (research participant); and (2) scores obtained by one occupational therapist rater administering the SOTOF to the same person on two separate occasions one day apart. The sample comprised of 32 occupational therapists and 37 older people. The sample comprised 54.1 percent females and 42.9 percent males, aged between 60 and 91 years. The majority (n = 21) of these patients had a primary diagnosis of stroke, 15 had dementia and 1 had a head injury. Several statistical analyses were undertaken; these included Percentage agreement, Pearson's Chi-square, Fisher's exact test, Phi Coefficient and Cohen's Kappa. Results indicated that both the average percentage agreement and approximate average Kappa values obtained on the SOTOF's sub-tests and Neuropsychological Checklist compared favourably with other Occupational Therapy standardised assessments. The SOTOF Screening Assessment appeared to have very good test-retest reliability (97.7 percent, Kappa approximate value of 0.92) and inter-rater reliability (97.5 percent, Kappa approximate value 0.94), and can be used as a reliable indication of gross motor, visual and cognitive functioning. The four SOTOF ADL Tasks have higher inter-rater reliability (90.3-93.8 percent, Kappa: 0.5-0.77) than test-rest reliability (89.5-91.6 percent, Kappa: 0.37-0.67). Examination of the reliability of the Neuropsychological Checklist found that the average percent agreement for test-retest reliability was 95.2 percent (approximate average Kappa value was 0.55) and inter-rater reliability was very similar at 95.2 percent (Kappa 0.54).

**Types of reliability:** This study focused on the evaluation of two types of reliability; test-retest reliability/consistency and inter-rater reliability/agreement. Reliability has been defined as the "consistency or stability of empirical indicators between raters or from one measurement to another ...it is the extent to which a measurement is free from random errors, ...it can be broadly defined as the consistency of a measurement" (Ottenbacher and Tomchek, 1993, p. 10). **Inter-rater reliability/agreement** refers to the "agreement between or among raters" (Ottenbacher and Tomchek, 1993, p. 11). Patients might be referred from one setting to another (e.g. ward to day hospital), or be re-referred after discharge. This can result in the need for a patient to be assessed by several different occupational therapists over a period of time. When this occurs it is important to gauge how likely a change in a patient's performance on a test is a result of a change in rater as opposed to a genuine change in the patient's level of ability. **Test-retest reliability** has been defined as the "correlation between the scores obtained by the same person on the two administrations of the test" (Anastasi, 1988, p. 116), and as the "consistency of an evaluation or test score over time" (Ottenbacher and Tomchek, 1993, p. 11). A similar methodology is used to evaluate both test-retest reliability and intra-rater reliability. Intra-rater reliability/agreement refers to "the consistency of judgements made by the same rater over a period of time" (Ottenbacher and Tomchek, 1993, p. 11). Frequently, an occupational therapist will wish to evaluate the effectiveness of a treatment programme by re-testing a patient on an assessment administered prior to treatment to see whether desired changes in function have occurred. It is, therefore, important that changes in a patient's performance on the test are not affected by the time interval or by the rater. A study was conducted to provide a measure of both the inter-rater and the test-retest / intra-rater reliability of the SOTOF.

**Methods for evaluating reliability**

Measurement of a subject on the SOTOF is interpreted in terms of the defined criterion behaviours which the person may or may not exhibit. If a subject is able to perform, and therefore pass, all the items in a task then that subject is considered to be independent for that task. The individual is not considered to have underlying neuropsychological deficits in any of the performance components which would impede his or her occupational performance in the Task's ADL domain. Criterion assessments usually have one of two main purposes: estimation of the domain score, i.e., the proportion of items in the domain which the subject can pass correctly; or mastery allocation. In mastery allocation the domain score is divided into a number of mutually exclusive mastery categories which are defined by cut scores. The observed test results are used to classify subjects into the mastery categories. "The most commonly cited example has one cut score and two categories, master and non-master" (Crocker and Algina, 1986). The concept of mastery allocation to one of two categories is applied to all the test items. The first phase standardized element of SOTOF uses a dichotomous, nominal scoring system; for each item there is an understanding of what the subject should be able to do in order to be classified in the master category which is labelled as 'able', conversely, failure to perform the item to this specified level results in the classification of non-master or 'unable'. The data produced from each

SOTOF item is therefore categorical and based on the judgement made by the therapist regarding the subject's ability or inability to perform the item. The evaluation of the reliability is concerned with the consistency or accuracy of the classification decisions made from the observation of the subject's performance. Analysis requires the application of a statistic to a two by two contingency table constructed for each item for (1) the first and second administration carried out by the same rater and (2) the same test administration scored by two different raters.

## Reliability study Research Questions

1. Is there correlation between scores obtained by two different occupational therapist raters scoring the same administration of the SOTOF to one patient? This question focused on the inter-rater reliability of the SOTOF.
2. Is there correlation between the scores obtained by one occupational therapist rater administering the SOTOF to the same patient on two separate occasions one day apart? This question sought to establish the test-retest and intra-rater reliability of the SOTOF.

## Methodology

The evaluation of reliability involved a combined sample obtained from two separate studies using the same methodology.

## First Reliability Study: Identifying and sampling the population

The sample population was drawn from two groups: qualified, hospital based occupational therapists working with older people with a diagnosis of stroke; and patients aged 60 years and over with a primary diagnosis of stroke. Patients with a recent onset of stroke are one of the target populations for the test. For this study, testing was to be undertaken no more than 12 weeks (3 months) from the onset of the stroke. The participants were drawn from a sample of occupational therapists recruited to the research as a result of a letter published in the British Journal of Occupational Therapy. The occupational therapists were contacted by telephone to take part in the reliability study. They were asked if they had a colleague who would be able to carry out the research with them. The therapists were working in hospitals within the United Kingdom.

## Procedure

Ethical approval was provided by the St George’s Hospital Medical School Ethics committee. Therapists agreeing to assist with the study were sent a packet comprising a letter, questionnaire, test manual and three sets of assessment forms. The letter gave details of the purpose of the study and the procedure. The questionnaire covered: (1) occupational therapists' details including year qualified, current clinical area, grade, experience working with elderly and stroke patients; (2) therapists' prior knowledge of the patient with an outline of previous intervention; (3) patients' details including their age, sex, primary / secondary diagnoses, and date of onset of stroke; and (4) assessment details including date, time and location of testing. The study was undertaken over a two-day period. On the first day, one of the occupational therapists administered the assessment to the patient and the second therapist observed the test administration. The two therapists were instructed to independently record their observations on the SOTOF Observational Task checklists and the Neuropsychological checklist. It was essential that there was no collaboration or conferring between the therapists. On the second day, therapists were instructed to have both tests administered by the same therapist, in the same test location and at the same time of day. They were told to record the testers' initials, date, time and location of testing for both test administrations on the questionnaire. The first author was available for clarification.

Other test developers have used video tapes of patients taking a test, completed test forms or drawings, and photographs of different arrangements of test items, to measure inter-rater reliability. These tapes, forms or photographs are scored by a number of different raters (e.g. Whiting et al, 1985). As the SOTOF involves the observation of four complete tasks, as well as the Screening assessment items, it would be difficult for a rater to gain a complete picture of the subject's performance from one frame or angle. It was impractical to film and edit videotape that had been shot from several angles. The participant does not complete any written or drawn items on the SOTOF and as the test involves the observation and evaluation of a person's action rather than an end product, (such as a those produced with block design or card sequencing items), photographing test items was also inappropriate. The SOTOF involves on-going clinical reasoning during the assessment. For example: decisions regarding the need for prompts or cues, such as the action 'on command' or 'when handed' object items; or the evaluation of language with the subsequent selection of different administration methods for some items dependent on whether the person has expressive language intact, such as the colour and object recognition items. Because of the nature of the test it was decided that people with varying levels of function should be tested and that the actual administration of the test should be observed by a second therapist. The two therapists (raters) agreed not to confer. However, it should be noted that the clinical reasoning element of the SOTOF is such that the observer could form opinion concerning the patient's function from the way the therapist gives certain test items. For example, if the therapist asks the patient to identify items though pointing rather than naming, the observer could determine that the patient has problems with expressive language.

## Second Reliability Study: Identifying and sampling the population

Additional participants were recruited from two hospitals in the south-east of England. Canterbury and Thanet Health Authority Ethical Committee approved the collection of data on the SOTOF for reliability, concurrent validity and normative studies, with participants who were clinically healthy people and/or had primary diagnoses of stroke, dementia, head injury or Parkinson's disease. The diagnostic categories for patient samples were increased at the request of occupational therapists that had taken part in earlier studies and felt that the SOTOF had relevance for an expanded population. Both in-patients and day-patients, under the care of local geriatricians and psychogeriatricians, were recruited for this study. One full time and three part-time occupational therapy research assistants were employed. Participants were recruited through referral from local consultant geriatricians and occupational therapists. The research assistants attended ward rounds and meetings in order to identify suitable patients for the study.

## Procedure

Once identified, the researcher visited potential participants on the ward or day hospital and provided an information leaflet outlining the project. A verbal explanation of the nature and purpose of the study was also provided at this stage. Potential participants were given time to discuss the project with their carers, relatives and/or friends and to read the information. When potential participants had visual or language deficits the leaflet was read out loud to them by the researcher or a member of their multidisciplinary team. Patients with stroke were to be tested on the wards and in the occupational therapy department of a local hospital; patients with dementia were to be tested at a psychogeriatric day hospital, on the wards of a second local hospital or at their own home. Prior to testing, the participant signed two copies of the consent form; one copy was attached to the patient's medical notes and the other was attached to their research records. The same testing procedure followed for the first reliability study was used for this study to allow the valid combination of the two samples for the statistical analysis.

## Description of sample and testing situation for the first study

Fourteen pairs of occupational therapists (n = 28) took part in this study and tested 14 participants with a primary diagnosis of stroke. One pair was not able to complete the assessment leaving 13 sets of completed data. The therapist **test administrators** had qualified between 1964 and 1991, and comprised of five basic grades, five senior II, two senior I, one head IV and one deputy head occupational therapist. Nearly half of the therapists were working in "geriatric" or "care of the elderly" settings (n = 6). The other therapists encountered elderly patients as part of their case load on neurology or medical and surgical wards. Therapists' experience with older patients ranged from less than 1 to 15 years: less than 1 (n = 2), 1 to 5 (n = 6), 6 to 10 (n = 1) and 11 to 15 (n = 3). The distribution for experience with stroke patients was similar: less than 1 (n = 2), 1 to 5 (n = 8), 6 to 10 (n = 1), 11 to 15 (n = 2). Eleven of the therapists had known the patient prior to the research. Pre-test intervention comprised of informal observation (n = 3), assessment (n = 4) or assessment and treatment (n = 4). Five therapists mentioned that they had previously administered an ADL assessment, two had carried out motor assessments, one had undertaken a sensory assessment, three patients had been cognitively assessed and three therapists had carried out perceptual assessments.

The therapist **observers** had qualified between 1967 and 1990, and comprised of three basic grades, six senior II, one senior I, two head IV, one head III and one occupational therapist of unspecified grade. Five of the therapists were working in geriatric or care of the elderly settings and the other therapists were based in medical, neurology, orthopaedics, rheumatology, outpatient and day hospital settings. Therapists' experience with older patients ranged from less than 1 to 15 years: less than 1 (n = 5), 1 to 5 (n = 7), 6 to 10 (n = 1) and 11 to 15 (n = 1). The distribution for experience with stroke patients ranged from less than one to 10 years: less than 1 (n = 3), 1 to 5 (n = 8), 6 to 10 (n = 2). Seven of the observing therapists had known the patient prior to the research. Intervention comprised of informal observation (n = 1), assessment (n = 1) or assessment and treatment (n = 4). Two therapists mentioned that they had previously administered an ADL assessment, one had carried out a motor assessment, one had undertaken a sensory assessment, one patient had been cognitively assessed and two therapists had carried out perceptual assessments.

Of the 14 participants who took part in this study, eight had Right Hemisphere Lesions resulting in left hemiplegia, four had Left Hemisphere Lesions resulting in right hemiplegia, and two had strokes of unspecified type. The time between onset of stroke and testing ranged up to three months: less than one month (n = 6), 1 to 2 months (n = 4), 2 to 3 months (n = 4). Secondary diagnoses varied with the most common being hypertension, diabetes or arthritis. Two participants had a history of previous stroke. The locations used for testing included: occupational therapy departments (1st test n = 6, retest n = 4); wards (1st test n = 3, retest n = 3); day hospitals (1st test n = 2, retest n = 2); rehabilitation units (1st test n = 1, retest n = 1); an activity unit (1st test n = 1, retest n = 1); a rehabilitation therapy area (1st test n = 1, retest n = 1); and a research room (1st test n = 0, retest n = 1).

### Description of sample and testing situation for the second study

The first author and three occupational therapy research assistants (one basic grade, one senior II, and one head occupational therapist) collected the data for the second study. Twenty-three participants were tested and the sample comprised of participants with the following primary diagnoses: stroke (n = 7); dementia (n = 15) and head injury (n = 1).

### Summary description of the combined sample

Data from the two studies was combined for the statistical data analysis. The overall sample was comprised of 32 occupational therapists (covering all grades from basic to head occupational therapist) and 37 participants (with primary diagnoses of: 21 stroke; 1 head injury; and 15 dementia). The participant sample contained 54.1 percent (n = 20) females and 45.9 percent (n = 17) males aged between 60 and 91 years (Mean 75.6, s.d. 8.2).

### Description of Statistical Analysis

At the time the studies were conducted (1991-1992) there was debate in the field of occupational therapy concerning the 'correct' statistic to use to estimate test-retest and inter-rater reliability. Ottenbacher and Tomchek (1993), reviewed 20 articles (from the American Journal of Occupational

Therapy and Physical Therapy), which reported reliability studies. Amongst the statistics discussed in their paper, those suitable for the type of data collected in this study were Kappa, chi-square, and percent agreement. Ottenbacher and Tomchek concluded that Kappa was one of "the preferred methods of computing reliability in applied environments" (p. 14); Kappa was preferred to percent agreement as it corrects for chance agreement. Discrepancies were found between the average Kappa values and the average percentage agreement indexes evaluated in their study; all the reliability coefficients in their study had a ceiling value of 1.00 or 100 percent, Kappa had an approximate average value of 0.5 compared to Percent agreement which had an approximate average of 0.75 (75 percent). It was, therefore, decided to compute several statistics for this study in order to compare the values obtained and examine whether the same items exhibit substantial differences in levels of reliability when reliability coefficients are calculated by the different statistical methods. All analyses were calculated using SPSS/PC+ software (Norusis, 1991). The statistical analyses undertaken for this study were: (1) Percentage agreement; (2) Pearson's chi-square, Fisher's exact test and Phi Coefficient; and (3) Cohen's Kappa. For all the analyses data, from the two test administrations or for the two raters, for each variable, was cross-tabulated in a two by two contingency table.

## Percentage agreement (P)

Percentage agreement (P) is an expression of the probability of a consistent decision (Crocker and Algina, 1986). P is the simplest measure of consistency for mastery decisions and can be defined as the proportion of people consistently classified as either master-master (able-able) or nonmaster-nonmaster (unable-unable) using two criterion referenced measurements. A new variable was constructed by assigning any subject who was consistently classified a value of one and inconsistently classified data a value of zero. P equaled the sum of these values divided by the maximum possible value of this sum (which can only be obtained if all decisions are consistent). P was then expressed as a percentage (Crocker and Algina, 1986). Some of the data in this study lacked variance; this resulted in the formation of one-by-one or two-by-one contingency tables. Addition statistics could not be calculated for these tables. As a result, percentage agreement was the only statistic that could be calculated for all test items, and was the value used to provide an estimate of the overall reliability of the SOTOF, the reliability of each of the items in the five sub-tests: i.e., Screening Assessment, Eating Task (Task 1), Washing Task (Task 2), Drinking Task (Task 3), and Dressing Task (Task 4) and the reliability of each of the items on the Neuropsychological Checklist. Detailed results of the analysis for each item can be found in Laver’s (1994) PhD thesis (Appendix 14 Tables 14.1 to 14.5). Two summary tables below (Tables 1 and 2), show the range of values and average value for each of the five sub-tests. The average percent agreement for test-retest reliability for the SOTOF was 91.8 percent (range 89.5-97.7 percent). The average percent agreement for inter-rater reliability was 93.1 percent (range 90.3-97.5 percent). The highest average values for both types of reliability were obtained for the Screening Assessment.

**Table 1: The Average percent agreement for test-retest reliability for the SOTOF**

| Sub-test | Range of % agreement across all sub-test items | Average % agreement for sub-test |
|---|---|---|
| Screening Assessment | 96.3% - 100% | 97.7% |
| Task 1 | 33.3% - 100% | 90.3% |
| Task 2 | 50.0% - 100% | 89.5% |
| Task 3 | 72.4% - 100% | 90.1% |
| Task 4 | 77.8% - 100% | 91.6% |
| | **Average % agreement for SOTOF** | **91.8%** |

**Table 2: The Average percent agreement for inter-rater reliability for the SOTOF**

| Sub-test | Range of % agreement across all sub-test items | Average % agreement for sub-test |
|---|---|---|
| Screening Assessment | 90.0% - 100% | 97.5% |
| Task 1 | 28.6% - 100% | 93.8% |
| Task 2 | 60.0% - 100% | 92.8% |
| Task 3 | 63.6% - 100% | 90.9% |
| Task 4 | 57.1% - 100% | 90.3% |
| | **Average % agreement for SOTOF** | **93.1%** |

An additional variable was constructed for the analysis of the reliability of the Neuropsychological Checklist. As the SOTOF is based on a progressive diagnostic clinical reasoning process, it was considered possible that therapists might reach the same decisions but from the observation of different tasks. It was, therefore, important to consider not just whether a specific deficit was recorded on the Neuropsychological Checklist under a specific sub-test heading, but whether raters identified the same deficits from the complete administration of the SOTOF. The new variable was constructed by giving a value of 1 (deficit present), to a participant whenever a deficit had been recorded in the Neuropsychological Checklist under the heading of at least one of the sub-tests and a value of 2 (deficit absent) when the deficit had not been recorded under any of the sub-test headings. Percentage agreement values for the Neuropsychological Checklist for each item can be found in Laver's (1994) PhD thesis (Appendix 14, Tables 14.6 to 14.11) and are summarised below in Tables 3 and 4. These tables show the range of values and average value for each of the five sub-test headings on the checklist (Screen, Tasks 1, 2, 3, and 4). The average percent agreement for test-retest reliability for the SOTOF Neuropsychological Checklist was 95.2 percent (range 92.4-97.6 percent). The average percent agreement for inter-rater reliability was 93.9 percent (range 90.5-96.6 percent). The combined test-retest percentage agreement for the SOTOF (sub-tests and

Neuropsychological Checklist) was 93.5 percent. The combined inter-rater value was 93.5 percent as well.

**Table 3: Average percent agreement for test-retest reliability for the SOTOF Neuropsychological Checklist**

| **Sub-test** | **Range of % agreement across all sub-test items** | **Average % agreement for sub-test** |
|---|---|---|
| Screening Assessment | 82.4% - 100% | 97.6% |
| Task 1 | 79.4% - 100% | 97.2% |
| Task 2 | 88.2% - 100% | 94.2% |
| Task 3 | 73.5% - 100% | 94.6% |
| Task 4 | 76.5% - 100% | 95.3% |
| Total | 67.6% - 100% | 92.4% |
| | **Average % agreement for SOTOF** | **95.2%** |

**Table 4: Average percent agreement for inter-rater reliability for the SOTOF Neuropsychological Checklist**

| **Sub-test** | **Range of % agreement across all sub-test items** | **Average % agreement for sub-test** |
|---|---|---|
| Screening Assessment | 87.5% - 100% | 96.6% |
| Task 1 | 79.2% - 100% | 94.2% |
| Task 2 | 79.2% - 100% | 93.5% |
| Task 3 | 79.2% - 100% | 93.6% |
| Task 4 | 83.3% - 100% | 94.6% |
| Total | 75% - 100% | 90.5% |
| | **Average % agreement for SOTOF** | **93.9%** |

### Chi-square, Fisher's exact test and Phi Coefficient

The null hypothesis for this analysis was that there was no relationship between the scores of the two raters or the scores from the two test administrations. **Pearson's chi-square** statistic was used to compare the observed score distributions to those that would be expected if the two variables (the two sets of test scores from inter-rater and test-retest studies), were independent. The reliable use of chi-square is dependent on sample size (Norusis, 1991; Spitznagel, 1991). Assumptions related to sample size with contingency tables are based on the expected frequencies (Portney and Watkins, 1993), whereby, "if some of the expected frequencies in a table are less than 5, the observed significance level based on the chi-square distribution may not be correct" (Norusis, 1991, p. 270). One way to counteract this problem is to collapse variables (Sigel and Castellan, 1988; Portney and Watkins, 1993), however, as the contingency tables were already based on dichotomous variables it was not possible to combine variables to increase the expected frequencies in the contingency table cells.

**Fisher's exact test** can be used to adjust chi-square to account for small expected frequencies and was calculated for this analysis. This test was used because it "evaluates the same hypothesis as the chi-square test, and it's suitable for tables having two rows and two columns for small expected frequencies" (Norusis, 1991, p. 270-271). Chi-square indicates if an association between variables is significant, the **Phi Coefficient** is used to express the degree of association between two nominal variables in a two-by-two table. The value of the Phi Coefficient ranges from -1.00 to +1.00 and can be interpreted as a correlation coefficient (Portney and Watkins, 1993). A significance level of 5% (< 0.05) was used to evaluate the significance of chi-square, Fisher's and Phi values. Values for these statistical computations were only available for a proportion of the sub-test and checklist items owing to a lack of variance. Summaries of results are shown below in tables 5 and 6. The full results can be found in greater detail in Laver’s (1994) PhD thesis (Appendix 14, Tables 14.1 to 14.11). Table 5 shows the total number of items that were significant at the <0.05 level (Pearson's Chi-square, Phi and Fisher's exact test) for test-retest and inter-rater reliability for the Screening Assessment, Eating Task (Task 1), Washing Task (Task 2), Drinking Task (Task 3), and Dressing Task (Task 4). Those items that were not significant at this level fall into three categories. First, it was not possible to calculate these statistics for all test items as some of the two-by-two contingency tables contained missing data values. Second, some items were significant at the <0.05 level for Pearson's Chi-square and Phi but not for Fisher's exact test (two sided probability). Third some items were not significant at the <0.05 for any of the statistical tests. A breakdown of the analysis for each test item is in Laver’s (1994) PhD thesis (Appendix 14, Tables 14.1 to 14.5).

**Table 5: Significance of inter-rater and test-reliability for the SOTOF Screening Assessment and Four ADL Tasks**

| **SOTOF component** | **Test-retest reliability: Number of significant items expressed as a fraction of the total number of items in that Task** | **Inter-rater reliability: Number of significant items expressed as a fraction of the total number of items in that Task** |
|---|---|---|
| Screening Assessment | 8/9 | 8/9 |
| Eating Task (Task 1) | 9/26 | 10/26 |
| Washing Task (Task 2) | 11/27 | 9/27 |
| Drinking Task (Task 3) | 6/28 | 10/28 |
| Dressing Task (Task 4) | 12/19 | 11/19 |

Results varied from item to item. All the items on the Screening Assessment were significantly related at the <0.05 level for both inter-rater and test-retest reliability, except for one item each that did not produce a two-by-two table. Only seven of the 26 items on Eating Task (Task 1: Eating from a Bowl using a Spoon) were not significantly related at the <0.05 level for test-retest reliability, and only three of the items on the Eating Task were not significantly related for inter-rater reliability. A similar distribution emerged for Washing Task (Task 2: Washing Hands in a Bowl): seven of the 27 items were not significantly related at the <0.05 level for test-retest reliability, and only two items were not significantly related for inter-rater reliability. For the Drinking Task (Task 3: Pouring and Drinking) seven of the 28 items for test-rest and three items for inter-rater reliability were not significantly related. In Dressing Task (Task 4: Putting on a Shirt), only two of the 19 items for test-retest and only one item for inter-rater reliability were not significantly related at the <0.05 level. Overall, the results indicated that the majority of items showed agreement across raters and, to a lesser extent, across time.

A pattern emerged for some types of items, from the four tasks, that were not significantly related at the <0.05 level. At least one of the "right / left discrimination" items was not significantly related for test-retest reliability on the first three tasks (Eating Task, Washing Task and Drinking Task). Patients rarely switch concepts of right and left completely but tend to exhibit general confusion in differentiating left from right. These items could have produced non-significant values because a deficit in right/left discrimination does not always result in a consistent response, but is more likely to appear as random performance with the subject sometimes placing the item correctly and sometimes giving an incorrect response.

The "recognition of objects" item was not significantly related for test-retest reliability in the Eating Task (Task 1), Washing task (Task 2) and Drinking Task (Task 3). The "describes use of objects" was also non-significant for test-retest reliability in three of the tasks (Eating Task, Drinking Task and Dressing Task). A possible explanation for these results could have been a learning effect if the subjects had been informed of the name and purpose of the objects by any of the raters during the first test administration. In clinical practice, therapists use assessment results as a starting point from which to educate patients. Raters in the second study had been trained by the researcher and did not offer such feedback. It was not possible to retrospectively examine whether raters from the first study had given feedback to patients following the first test administration. Further research would be required to clarify this point.

The 'when handed' objects items were not significantly related at the <0.05 level for inter-rater reliability for all four tasks. This could have resulted from some ambiguity regarding both the administration and scoring of these items. This ambiguity came to light during the norming and was clarified in the original SOTOF test manual (Laver and Powell, 1995)..

Other items that were not significantly related at the <0.05 level, appeared to be randomly distributed across tasks or only occurred in one of the four tasks. The test-retest reliability of the colour recognition items, for example, was significantly related for all but the Dressing Task (Task 4). The colours on the other three tasks could have been easier to perceive owing to the size of the objects and because brighter primary colours were used (the button used for the second study was dark blue). This problem might be solved by increasing the size of the button used and changing to an easily perceived colour, such as yellow or red. This would also address the problem of using dark colours from the blue/green end of the spectrum which are more difficult for older people to perceive owing to primary ageing which causes yellowing of the retina. There could have been a learning effect on this item if any of the raters had corrected the patient and informed them of the colour of the button during the first test administration.

The Screening Assessment is used to evaluate whether the person is functioning at the baseline level defined in the criteria for the administration of the SOTOF. Patients, therefore, should have passed the majority of the Screening Tasks if they had been entered in the rest of the study. Because of this high pass rate many of the deficits under the Screen heading of the Neuropsychological checklist lacked variance and statistics could not be computed for a large proportion of these items. (Percentage agreement for these items was very high ranging from 82.4% to 100% with an average of 97.6%). All of these items were significantly related for inter-rater reliability indicating considerable agreement among test administrators. All but two items were significantly related for test-retest reliability, these were expressive language and hearing acuity. Both these functions would not have been expected to alter in stroke patients during such a short space of time. The non-significant value obtained for the hearing acuity item is more likely to be the result of random errors; possible explanations include changes in the level of background noise in the testing environments or the failure of the participant to use a hearing aid (if required), during one of the two test administrations.

Summaries of the results for items on the Neuropsychological Checklist are provided in Table 6. The full detailed results can be found in Laver's (1994) PhD thesis (in Appendix 14, Tables 14.6 to 14.11). Table 6 shows the total number of items that were significant at the <0.05 level (Pearson's Chi-square, Phi and Fisher's exact test) for test-retest and inter-rater reliability for each Neuropsychological deficit under the five Checklist headings (Screen, Task 1, Task 2, Task 3, and Task 4) and the constructed "Total" variable. All values are presented as a fraction of the total number of Neuropsychological Checklist items for each deficit (i.e. out of a total of six items per deficit). Those items that were not significant at this level fall into the same three categories described above. Many of the Neuropsychological Checklist items did not produce statistical values owing to lack of variance; it should be noted that the majority of these items had a percentage agreement of 100%. Some items were significantly related at the level <0.05 level but only for Pearson's Chi-square and Phi, not for Fisher's exact test.

### Table 6: Significance of Reliability of Neuropsychological Checklist Items

| Deficit | Test retest: Number of significant items (maximum = 6) | Inter rater: Number of significant items (maximum = 6) |
|---|---|---|
| Language: comprehension | 5 | 2 |
| Language : expression | 3 | 4 |
| Hearing : acuity | 3 | 1 |
| Hearing : auditory agnosia | 0 | 0 |
| Cognition : orientation | 1 | 1 |
| Cognition : attention | 5 | 0 |
| Cognition : short term memory | 1 | 2 |
| Cognition : long term memory | 1 | 0 |
| Motor : abnormal tone | 6 | 6 |
| Sensation : proprioception | 6 | 6 |
| Sensation : tactile discrimination | 2 | 0 |
| Vision : acuity | 0 | 0 |
| Vision : Visual attention | 0 | 0 |
| Vision : visual scanning | 0 | 0 |
| Vision : visual field loss | 0 | 0 |
| Vision : visual neglect | 2 | 0 |
| Agnosia : visual spatial | 0 | 0 |
| Agnosia : visual object | 1 | 1 |
| Agnosia : colour agnosia | 0 | 0 |
| Agnosia : tactile agnosia | 5 | 3 |
| Apraxia : constructional | 2 | 0 |
| Apraxia : dressing apraxia | 2 | 2 |
| Apraxia : Motor apraxia | 2 | 0 |
| Apraxia : ideomotor apraxia | 2 | 0 |
| Apraxia : ideational apraxia | 2 | 0 |
| Body Scheme : somatognosia | 0 | 0 |
| Body Scheme : unilateral neglect | 3 | 4 |
| Body Scheme : anosognosia | 0 | 0 |
| Body Scheme : right / left discrimination | 1 | 3 |
| Spatial Relations : figure ground | 3 | 0 |
| Spatial Relations : position in space | 3 | 1 |
| Spatial Relations : form constancy | 0 | 0 |
| Spatial Relations : spatial relations | 3 | 1 |
| Spatial Relations : depth perception | 0 | 0 |
| Spatial Relations : distance perception | 0 | 0 |
| Perseveration : | 0 | 0 |

A pattern emerged for some of the non-significant items, for example, the 'Sensation: tactile discrimination' was inconsistently recorded by raters across all four Task headings and as examined through the constructed 'Total' variable. Examination of the Task observational checklist assessment forms showed an inconsistency between raters regarding the scoring for the 'identifies object through touch-left hand' item, especially when the subject had previously identified the object with his/her right hand. The 'Agnosia: tactile agnosia' item, which is also identified through the performance of these 'identification through touch' items, were inconsistently recorded for both test-retest and inter-rater reliability.

The 'Cognition: short term memory' item was not significantly related for both test-retest and inter-rater reliability under the Eating Task (Task 1), Washing Task (Task 2) and Drinking Task (Task 3), Neuropsychological checklist headings. Both types of reliability, however, were significantly related when examined through the constructed 'Total' variable for this deficit. This suggests that the short term memory deficit is identified consistently overall by the test administrators, but does not manifest during any one specific Task performance. A similar pattern also emerged for 'Cognition: attention' which had significant values for both types of reliability for the Total variable despite non-significant inter-rater reliability values for the Eating Task (Task 1) and Drinking Task (Task 3), and a non-significant test-retest value for Washing Task (Task 2).

Other deficits that were not significantly related at the <0.05 level for some of the tasks but which were consistently recorded over the whole checklist as indicated by significant 'Total' values were: 'Language: expression'; 'Agnosia: visual object agnosia'; 'Apraxia: ideomotor apraxia'; 'Apraxia: ideational apraxia'; 'Body scheme: right/left discrimination'; and 'Spatial relations: spatial relations'. Conversely only three deficits produced non-significant values for the Total variable: 'Vision: visual attention' was non-significant for test-retest reliability, 'Agnosia: visual spatial' and 'Spatial relations: figure ground discrimination' were non-significant for inter-rater reliability. Deficits which had some items that were not significantly related for some Task headings and for the Total variable were: 'Language: comprehension'; 'Hearing: acuity'; 'Cognition: long term memory'; 'Vision: visual scanning'; 'Vision: visual field loss'; 'Apraxia: constructional apraxia'; and 'Perseveration'. The number of items that were not significantly related at the <0.05 level for these deficits ranged from two to six.

## Cohen's Kappa (K)

Cohen's Kappa (K) is a measure of agreement which has "been proposed for categorical variables [and] can be applied to an arbitrary number of raters" (Siegel and Castellan, 1988, p. 284). Kappa provides a transformation of P to a new scale in which the points 0 and 1 are interpretable:

> "where Pc is the chance probability of a consistent decision... that is, the probability for the hypothetical situation in which the scores on the two forms are statistically independent. Statistical independence of test

> scores implies that decisions are statistically independent. The coefficient Pc is sometimes referred to as the *chance consistency*... chance consistency can be viewed as a baseline for judging the actual amount of consistency observed for the two forms [administrations of the test]. Thus K may be interpreted as the increase in decision consistency that tests provide over chance expressed as a proportion of the maximum possible increase over chance consistency"
>
> (Crocker and Algina, 1986, p. 200-201)

Coefficient K is 0 when there is no increase and 1.0 when there is maximal increase. A value of 0 does not mean that decisions are so inconsistent as to render the item worthless, but that the decisions are no more consistent than decisions based on statistically independent scores. This consistency could still be substantial (a minimum of 50% (0.5) for exchangeable test forms). A value of 1 indicates that decisions are as consistent as those based on perfectly statistically dependent scores (Crocker and Algina, 1986; Siegel and Castellan, 1988; Norusis, 1990). "The coefficient K can assume negative values...which corresponds to the situation in which there is an inverse relationship between the scores on the two forms" (Crocker and Algina, 1986, p. 201). Kappa treats all inconsistent classifications as equally serious. As the SOTOF does not use a continuous scoring system or scale, statistics which evaluate the magnitude of the discrepancy of a misclassification in judging reliability of decisions were not relevant.

SPSS/PC+ was used to compute Cohen's Kappa with asymptotic standard error (ASE1) and the t statistic value. "The test of the null hypothesis that kappa is 0 can be based on the t statistic... The t value is the ratio of the value of kappa to its asymptotic standard error when the null hypothesis is true. [N.B.] the asymptotic standard error on the [SPSS/PC+] output does not assume that the true value is 0" (Norusis, 1990, p. 136-137). Full results for Kappa, ASE1, and t values for each of the SOTOF sub-test items and the Neuropsychological Checklist items can be found in Laver’s (1994) PhD thesis (Appendix 14, Tables 14.12 to 14.22). Kappa values were only available for a proportion of the test and checklist items owing to lack of variance. Only one item in the entire test (test-retest reliability Washing Task 'continues action unnecessarily') obtained a value of zero which indicated that decisions were no more consistent than decisions based on statistically independent scores. The scoring of this item was identified as ambiguous during the Norming Study and was clarified in the original SOTOF test manual (Laver and Powell, 1995). Nine sub-test items obtained a value of one for test-retest reliability indicating that decisions were as consistent as those based on perfect statistically dependent scores. Fourteen sub-test items also obtained a value of one for inter-rater reliability. On the Neuropsychological Checklist 15 items had a value of one for test-retest reliability and nine for inter-rater reliability. It was impossible to obtain Kappa values for all test items, average Kappa values could only be calculated from a proportion of the items and should, therefore, be viewed as approximate values. Average Kappa values are shown in Tables 7 to 8.

**Table 7: approximate average Kappa values for the Screening Test and four ADL Tasks for test-retest reliability**

| Sub-test | Number of SOTOF items that kappa could be calculated for | Range of Kappa values across sub-test items | Average Kappa value for sub-test |
|---|---|---|---|
| Screening Assessment | 10 / 11 | 0.78 - 1 | 0.92 |
| Task 1 | 17 / 26 | -0.04 - 0.9 | 0.47 |
| Task 2 | 19 / 27 | -0.07 - 0.77 | 0.38 |
| Task 3 | 12 / 28 | -0.09 - 0.66 | 0.37 |
| Task 4 | 15 / 19 | -0.07 - 1 | 0.67 |
| **Average Kappa value for SOTOF** | 73 / 111 | -0.09 - 1 | 0.56 |

**Table 8: approximate average Kappa values for the Screening Test and four ADL Tasks for inter-rater reliability**

| Sub-test | Number of SOTOF items that kappa could be calculated for | Range of Kappa values across sub-test items | Average Kappa value for sub-test |
|---|---|---|---|
| Screening Assessment | 8 / 11 | 0.65 - 1 | 0.94 |
| Task 1 | 10 / 26 | -0.4 - 1 | 0.77 |
| Task 2 | 7 / 27 | 0.23 - 1 | 0.5 |
| Task 3 | 8 / 28 | 0.25 - 1 | 0.61 |
| Task 4 | 12 / 19 | 0.4 - 1 | 0.75 |
| **Average Kappa value for SOTOF** | 73 / 111 | -0.4 - 1 | 0.71 |

The approximate average Kappa values for the Screening Test and four ADL Tasks ranged from 0.37 to 0.92 (average 0.56) for test-retest reliability and from 0.5 to 0.94 (average 0.71) for inter-rater reliability. The overall average Kappa value for test-retest reliability for the SOTOF was calculated from values available for 53.2% of items and was 0.56. The overall average Kappa value for inter-rater reliability for the SOTOF was calculated from values available for 40.7% of items and was 0.63. These values are slightly above the average Kappa value (0.5), reported by Ottenbacher and Tomcheck (1993) in their evaluation of reliability analysis in therapeutic research.

**Table 9: approximate average Kappa values for the Neuropsychological checklist for test-retest reliability**

| Checklist Sub-test Heading | Number of SOTOF items that kappa could be calculated for | Range of Kappa values across sub-test items | Average Kappa value for sub-test |
|---|---|---|---|
| Screening Assessment | 7 / 36 | 0.21 - 1 | 0.63 |
| Task 1 | 18 / 36 | -0.04 - 1 | 0.56 |
| Task 2 | 18 / 36 | -0.05 - 0.67 | 0.44 |
| Task 3 | 15 / 36 | -0.06 - 0.84 | 0.47 |
| Task 4 | 16 / 36 | -0.05 - 1 | 0.61 |
| Total | 27 / 36 | -0.04 - 1 | 0.59 |
| **Average Kappa value for SOTOF** | / 216 | -0.06 - 1 | 0.55 |

**Table 10: approximate average Kappa values for the Neuropsychological checklist for inter-rater reliability**

| Checklist Sub-test Heading | Number of SOTOF items that kappa could be calculated for | Range of Kappa values across sub-test items | Average Kappa value for sub-test |
|---|---|---|---|
| Screening Assessment | 6 / 36 | 0.47 - 1 | 0.8 |
| Task 1 | 14 / 36 | -0.09 - 1 | 0.52 |
| Task 2 | 13 / 36 | -0.06 - 1 | 0.52 |
| Task 3 | 13 / 36 | -0.11 - 1 | 0.47 |
| Task 4 | 17 / 36 | -0.04 - 1 | 0.5 |
| Total | 25 / 36 | -0.07 - 1 | 0.44 |
| **Average Kappa value for SOTOF** | 88 / 216 | -0.09 - 1 | 0.54 |

### Comparison of the values obtained by each of the statistical analyses

Summary tables for the three analyses were constructed for the items on the Neuropsychological Checklist to examine the discrepancy of reliability values obtained through each of the statistical methods and can be found in Laver's (1994) PhD thesis (Appendix 14, Tables 14.23 to 14.26). Comparison of percentage agreement and Kappa values obtained in this study supported the finding by Ottenbacher and Tomchek (1993) that percentage agreement values were consistently higher than kappa values. Average percentage agreement for test-retest reliability of the SOTOF (calculated from values for all items) was 0.94 (93.5%), compared with an approximate (calculated from values available for only 53.2 percent of items) average Kappa value of 0.56. Average percentage agreement for inter-rater reliability of the SOTOF (calculated from values for all items), was also 0.94 (93.5%) compared with an approximate (calculated from values available for only 40.7 percent of items), average Kappa value of 0.63. Comparison of Kappa values with the significance level of values obtained by Chi-square, Fisher's and Phi showed that items with Kappa values of 0.5 and above were usually significant (at the <0.05 level) for these other analyses. Items with Kappa values between 0.34 and 0.65 were significant for Chi-square and Phi but did not always produce significant values for Fisher's exact test. Items with Kappa values less than 0.34 usually had non-significant values for the three other statistical analyses.

## Conclusions

The use of Cohen's Kappa, Chi-square (adjusted for small sample sizes where necessary) and Phi Coefficient produce more conservative estimates of reliability than Percentage agreement and are, therefore, preferred methods of analysis. The Kappa value is easy to interpret and gives the advantage of accounting for chance agreement; the results of this study supported Ottenbacher and Tomchek's (1993), recommendation of Kappa as a preferred method of computing reliability in applied therapeutic research. Unfortunately, a lack of variance in some of the data meant that Kappa could not be calculated for all the SOTOF test items. The average Kappa values are, therefore, only approximations of the overall reliability. It was necessary to rely on Percentage Agreement values, however, they should be treated with some caution as they may give an over positive image of the test's reliability.

The Screening Assessment appears to have very good test-retest (97.7 percent, Kappa approximate value of 0.92), and inter-rater reliability (97.5 percent, Kappa approximate value 0.94), and can be used as a reliable indication of gross motor, visual and cognitive functioning. The four ADL Tasks have higher inter-rater reliability (90.3-93.8 percent, Kappa: 0.5-0.77) than test-rest reliability (89.5-91.6 percent, Kappa: 0.37-0.67). This could be the result of genuine fluctuations in subjects' performance over the two administrations of the test. The research assistants who conducted the testing for the second study noted what they considered to be genuine changes in the performance of some participants with dementia from one test administration to another. A few of the occupational therapists who conducted the first study noted changes in the performance of some of their stroke patient subjects. This was partly the result of participants responding to therapists' corrections during the first test administration (e.g. learning a hemiplegic dressing method shown during the first test enabled independent dressing in the second test), and to perceived changes in function from one day to the other. Patients in the early stages following stroke can make spontaneous recovery. Furthermore the rationale behind practice of ADL tasks in occupational therapy is based on the belief that repetition of tasks aids the return of function. The fact that the re-test was a repetition of task performance could have also resulted in some slight increase in functional performance.

Both the average percent agreement and Kappa values for the SOTOF are higher than the average of the values reported in the reliability studies evaluated by Ottenbacher and Tomchek (1993): SOTOF's test-retest average of 91.8 percent and inter-rater average of 93.1 percent were higher than the average values for these 20 studies which was approximately 75 percent; average Kappa values of 0.56 for test-retest and 0.71 for inter-rater reliability were also higher than the 0.5. average value reported for these studies.

As the Neuropsychological Checklist score is based on diagnostic reasoning and requires rater judgement, it was anticipated that its reliability would be less than the SOTOF Tasks, and lower than other Neuropsychological Assessments. However, the average percent agreement for test-retest reliability was 95.2 percent and the approximate average Kappa value was 0.55. Inter-rater reliability was very similar at 95.2 percent / 0.54. These figures are encouraging, particularly when the ranges of experience of the clinicians used in this study are taken into consideration.

The average percentage agreement and approximate average Kappa values obtained on the SOTOF's sub-tests and Neuropsychological Checklist compared favourably to other occupational therapy standardised assessments available at the time of the study (early 1990s). The SOTOF values were particularly encouraging in light of the fact that the test involves a major component of rater judgement (therapist’s clinical reasoning). This supported the supposition that observation of a patient's performance in ADL tasks can provide as reliable a picture of neuropsychological deficit as the more formal psychological test batteries currently in use.

## References:

Crocker, L. & Algina, J. (1986). Introduction to Classical and Modern Test Theory. London: Holt, Rinehart and Winston, Inc.

Laver, A. J. (1994). *The development of the Structured Observational Test of Function (SOTOF).* PhD thesis. Guildford: Department of Psychology, University of Surrey. (Available from the College of Occupational Therapists library in London, UK)

Laver, A. J. & Powell, G. E. (1995). *The Structured Observational Test of Function (SOTOF).* Windsor: NFER-NELSON.

Norusis, M. J. (1990). *SPSS Base System Users Guide.* Chicago: SPSS Inc.

Ottenbacher, K. J. & Tomchek, S. D. (1993). Reliability Analysis in Therapeutic Research: Practice and Procedures. *American Journal of Occupational Therapy*, *47,* 10 - 16.

Portney, L. G. & Watkins, M. P. (1993). *Foundations of Clinical Research - Applications to Practice.* Connecticut: Appleton & Lange.

Spitznagel, E. (1991). Statistics - Course Notes for Math 320 - Academic Year 1991 - 1992. St Louis: Department of Mathematics, Washington University.