

Neuropsychological detection of cognitive impairment: Inter-rater agreement and factors affecting clinical decision-making

HOLLY A. TUOKKO, GILLIAN GABRIEL, AND THE CSHA NEUROPSYCHOLOGY
WORKING GROUP*

University of Victoria, Victoria, British Columbia, Canada

(RECEIVED January 13, 2005; FINAL REVISION September, 22; 2005; ACCEPTED September 25, 2005)

Abstract

The agreement between neuropsychologists identifying cognitive impairment (CI) in older adults was examined, as were factors influencing the classification process. Twenty four neuropsychologists in 18 study centers classified cases with or without CI after reviewing neuropsychological findings and other relevant information. All cases were participants in the third wave of the Canadian Study of Health and Aging, a study of CI in later life. For 117 randomly selected cases, a second neuropsychologist reviewed the same material and reclassified the cases. Cases given the same (concordant) or different (discordant) classifications were compared with respect to patient and rater characteristics. The inter-rater agreement was moderate (77.7% agreement, kappa = .49). On all measures of cognitive functioning, the concordant group without impairment obtained a higher mean score than the discordant group, and the discordant group obtained a higher mean score than the concordant group with impairment. For 5 out of 8 cognitive measures, the concordant group with impairment differed from the concordant group without impairment and the discordant group, but the latter two groups did not differ significantly. The findings are comparable to others in the field and highlight the need for neuropsychologists to further clarify procedures for identifying subtle, or mild, forms of cognitive impairment. (*JINS*, 2006, 12, 72–79.)

Keywords: Aged, Aged 80 and over, Classification, Reliability, Cognition disorders, Clinical psychology

INTRODUCTION

The identification of cognitive impairment has long been recognized as critically important to the health and well

being of older adults. Cognitive impairment (CI), regardless of its severity, may signal the need for a medical evaluation to determine the etiology of the impairment and inform the clinician regarding possible treatment options. In addition, it is becoming recognized that mild cognitive impairment seen in older adults may evolve to dementia in a significant proportion of persons (e.g., Tuokko & Frerichs, 2000) and that many cognitively impaired older persons, with or without dementia, require community-based support services if they are to remain in their own homes (Shapiro & Tate, 1997).

Cognitive impairment, as seen in older adults, is a heterogeneous classification and its identification relies heavily on a broad understanding of brain-behavior relations across the lifespan. It is well known that many different underlying disorders may result in impaired cognitive functioning and that the prevalence of disorders affecting cognition increases with age (e.g., Canadian Study of Health and Aging Working Group, 1994). Given the complexity

Reprint requests to: Holly A. Tuokko, Centre on Aging, Sedgewick Building Room A104, University of Victoria, PO Box 1700 STN CSC, Victoria, B.C., Canada V8W 2Y2. E-mail: htuokko@uvic.ca

*The following persons actively participated in the CSHA Neuropsychology Working Group: Ian McDowell, University of Ottawa, Ottawa; John D. Fisk, Dalhousie University, Halifax; Mary C. Tierney, Sunnybrook & Women's College Health Sciences Centre, University of Toronto, Toronto; Margaret Crossley, University of Saskatchewan, Saskatoon; Martine Simard, University of Toronto, Toronto; Nancy Fisher, University of Alberta Hospital, Edmonton; Betsy Kristjansson, University of Ottawa, Ottawa; Robert Laforce, University of New Brunswick, Saint John; Bernadette Ska, Institut Universitaire de Geriatrie de Montreal and Université de Montréal, Montreal; Gary Snow, Private Practice, Toronto; Jim Woodrow, Memorial University of Newfoundland & Health Care Corporation of St. John's, St. John's; Johanne Bernier, University of Sherbrooke, Sherbrooke; Andree Tellier, Ottawa Hospital, Ottawa; Lori Della Malva, The Rehabilitation Centre, Ottawa; Lisa Partlo, Foothills Medical Centre, Calgary.

and importance of this task, it is vital to ascertain whether clinicians draw similar conclusions when provided with the same information. That is, the inter-rater agreement when identifying cognitive impairment is of the utmost importance for neuropsychologists.

When identifying cognitive impairment in older adults, a number of factors may need to be taken into consideration. For example, O'Connor et al. (1996) cited factors contributing to discrepant clinical judgments, such as participants' poor vision or deafness, and rater confidence. In other fields, clinician experience and training have been found to influence inter-rater reliability (e.g., Steinhausen & Erdin, 1991; Ballantyne et al., 1995; Brooks & Thomas, 1995). Thus both the characteristics of the participants (e.g., sensory impairment, health status) and the raters (e.g., experience, training, confidence levels) may influence clinical decision-making.

A number of studies have examined inter-rater agreement in the context of specific disorders affecting cognition (e.g., Alzheimer's disease, Kukull et al., 1990; Lopez et al., 1990; Hogervorst et al., 2000; the clinical identification of dementia using DSM-III-R criteria, Baldereschi et al., 1994; Solari et al., 1994; Graham et al., 1996; O'Connor et al., 1996), but few have examined the agreement between raters in the detection of CI more generally. Only one study, that included older adults, specifically examined the inter-rater agreement among neuropsychologists (5 clinical neuropsychologists from 4 medical centers) when identifying CI (White et al., 2002). Two hundred and fifty-one cases were selected from the participating centers after excluding those with recent stroke, neurosurgery, head injury, legal blindness, alcohol or drug abuse, or severe deafness. Each neuropsychologist reviewed the medical, neuropsychological, and interview data for cases from his/her own site, while the members of an external review panel, 4 senior neuropsychologists, each reviewed 25% of the cases. The agreement between the individual neuropsychologists' and the review panel for the classification of CI *versus* no CI was moderate (κ coefficient = .48).

We examined the reliability with which CI was identified by neuropsychologists in a large, national, epidemiological study of cognitive functioning in late life. Our goals were to assess the inter-rater agreement of clinician-based neuropsychological classification of CI, and to identify factors that influenced the decision-making process. We identified cases where the neuropsychologists' classifications were concordant or discordant and compared the resulting groups on patient and rater characteristics. It was anticipated that there would be two types of concordant cases: those who were clearly not impaired and those who were clearly impaired (O'Connor et al., 1996). Further, it was anticipated that discordant cases would fall between the concordant not impaired and concordant impaired cases on measures of cognitive functioning (i.e., cognitive screening and neuropsychological measures) and not be clearly distinguishable from them (Graham et al., 1997). In relation to the concordant not impaired, the concordant impaired were

expected to be older, have less education, be female, have lower scores on measures of cognition, depression, and premorbid IQ, as all of these characteristics have been shown to differ between those with and without dementia (e.g., Jorm et al., 1998; Stern et al., 1994). In addition, given that the raters received the same information and the population under study was fairly homogeneous (e.g., age), it was anticipated that rater characteristics may also influence the decision-making process.

METHOD

Overall CSHA Study Design and Procedure

Cases for inclusion in this study were drawn from participants in the third wave of the Canadian Study of Health and Aging (CSHA-3). The CSHA is a national, longitudinal investigation of CI and dementia in which data were gathered at 18 study sites across Canada at five-year intervals (in 1990, 1996, 2001). Each participant underwent the same assessment in either English or French. Of the Original CSHA cohort established in 1990, 3,424 survivors took part in the CSHA-3 study, and 1,484 of these took part in the neuropsychological examination.

The participants in the CSHA-3 neuropsychological assessment were: (1) those who received a CSHA-1 or CSHA-2 consensus diagnosis of no cognitive impairment (NCI) or cognitive impairment with no dementia (CIND); and (2) participants who screened negative for cognitive impairment at CSHA-2 and, during the CSHA-3 screening examination, obtained a score falling above 49 but below 90 on the Modified Mini-Mental State Examination (3MS; Teng & Chui, 1987).

Neuropsychological Assessment

The neuropsychological test battery used in CSHA-3 was a subset of 8 of the measures used in CSHA-1 and 2 (Tuokko et al., 1995), and included the following: A Canadian version of the Wechsler Memory Scale Information subtest (Wechsler, 1975) and a modified version of the Buschke's Cued Recall paradigm (BCR; Tuokko et al., 1991) were used to assess memory; the Digit Symbol subtest of the Wechsler Adult Intelligence Scale-Revised (WAIS-R, Wechsler, 1981) and short forms (Satz & Mogel, 1962) of the WAIS-R Similarities and Block Design subtests were used to assess abstract reasoning and construction. Language skills were assessed using measures of the generation of words in response to letter (Controlled Oral Word Association Test; Spreen & Benton, 1977) and semantic cuing (animal names; Rosen, 1980), and a version of the Boston Naming Test using the 30 even items (Fisher et al., 1999). The Reading subtest from Wide Range Achievement Test-3 (Wilkinson, 1993) was administered to assess premorbid intelligence (note: an equivalent measure was not available for the sample assessed in French). This battery was administered and scored by trained psychometricians who also recorded

whether or not the participants exhibited problems that impeded testing in each of the following areas: hearing, vision, fatigability, inattention, perseveration, impulsivity, social impropriety, tangentiality, physical impairments, and facility with testing. The psychometrician also rated participant cooperativeness from 1 (Excellent) to 5 (Poor) and participant facility with language from 1 (Completely fluent) to 5 (Major difficulty). *T* scores for the neuropsychological tests were determined using the age, education, and gender corrected norms developed by Tuokko & Woodward (1996) in an English-speaking sample for the following measures derived from this battery: BCR Trial 1 Free Recall, BCR Total Retrieval (Free Recall Trial 1 + Trial 2 + Trial 3), WAIS-R Digit Symbol, WAIS-R Block Design, WAIS-R Similarities, Animal Naming, and Verbal Fluency. Fisher et al.'s (1999) norms for the Boston Naming Test were used to interpret performance on this measure. *T* scores could not be calculated for the French-speaking sample, but, where possible, appropriate norms were applied.

NCI and CI Classification: Decision-Making Process

The neuropsychological assessments were evaluated by a neuropsychologist who had access to all information obtained from the participant (i.e., self-reported) during the screening interview: the participant's marital status, vision, hearing, living situation, social support, activities of daily living, health status, health conditions, leisure and physical activities, income, and health care and end-of-life preferences. In addition, the neuropsychologist had access to the participant's scores on the following measures administered as part of the screening interview: the 3MS, the Short Happiness and Affect Research Protocol (Stones et al., 1996), and the Center for Epidemiologic Studies-Depression scale (Andresen et al., 1994).

The neuropsychologists were asked to make a clinical judgment as to whether or not CI was evident; they were specifically requested to include even very mild forms of CI (e.g., 1–1.5 *SD* below mean on neuropsychological tests). The following guidelines, provided to the neuropsychologists, were chosen to reflect the way in which cognitive impairment of insufficient magnitude to warrant a diagnosis of dementia has been described in the literature (e.g., Petersen et al., 1999). Guideline 1: Cognitively Impaired if score on any test is 1.5 *SD* or more below the mean, or Guideline 2: Cognitively Impaired if score on more than 1 test is 1.0 *SD* or more below the mean. Their use was intended to encourage clinicians to include people with mild CI in addition to those more severely impaired. The neuropsychologists were also asked to rate their confidence in their classification on a four-point scale (very, moderately, somewhat, not at all confident). The cases identified with CI by the neuropsychologists then proceeded to a full medical assessment after which a consensus diagnosis concerning the underlying pathology of the CI was determined.

The goal of this approach to identifying cases for full medical appraisal in the context of a large, epidemiological study was to reduce costs, as reliance on a strictly psychometric approach has been associated with high false positive rates when applied with cognitive screening measures like the Mini-Mental State Examination (O'Connor et al., 1991). As this was not a strictly controlled experiment, but a naturalistic clinical research study, it was anticipated that the neuropsychologists might not adhere rigidly to the guidelines provided or be unduly tied to test performance criteria, but that they would exercise reasoned clinical judgment in making their classifications. For example, a participant might have met Guideline 1 on one measure but the neuropsychologists might conclude that no cognitive impairment was present, as (following the definition of a standard deviation) a substantial minority of the normal population would be expected to fall in this range on a single measure. Conversely, a participant who did not meet Guideline 1 or 2 may have been identified as cognitively impaired even though his/her performance fell above the criteria but below their estimated premorbid level of functioning.

The Present Study of Inter-Rater Agreement

Participants

After the CSHA-3 data collection was finished, 117 cases were randomly selected for rereview by a second neuropsychologist from a different study center. Of the 117 CSHA-3 participants who comprised the reliability sample, ages ranged from 75 to 100 years, with a mean of 84.5 years, *SD* = 5.7. Seventeen of these persons were assessed in French and 100 were assessed in English.

The Original classifications of CI/no CI were provided for the 117 cases by 24 neuropsychologists recruited by the study centers involved in CSHA-3. Nine of these 24 neuropsychologists then volunteered to provide CI/no CI reclassifications for cases from other centers, forming the Reliability classifications; these neuropsychologists did not reclassify any cases for which they provided the Original classification. Four additional neuropsychologists, involved in the CSHA-3 but who did not provide Original classifications, volunteered to provide Reliability classifications. Hence, a total of 28 neuropsychologists (15 Original only, 9 Original and Reliability, 4 Reliability only) were involved in classifying the 117 cases. All neuropsychologists were surveyed with respect to their degree status, specialty designations, years of experience, and familiarity with each neuropsychological test, rated as "not at all familiar" (0) to "very familiar" (10).

Of the neuropsychologists, 24 held doctoral degrees in psychology and 4 held Master of Arts or Science degrees. All were licensed to practice as clinical psychologists except for one who was acting under the supervision of a registered psychologist (note that registration with a Master's degree is accepted practice in some provinces in Canada).

Eight of the neuropsychologists who took part were involved only with CSHA-3, whereas 13 took part in 2 waves of the CSHA, and 7 took part in all 3 waves of the study. All neuropsychologists had experience in assessing elderly people and only 3 indicated that assessment of older persons comprised less than 50% of their clinical caseloads. Most neuropsychologists were very familiar with the neuropsychological measures; only 2 had familiarity ratings below 6 on more than 1 measure. The neuropsychologists were least familiar with Buschke's cued recall paradigm for memory assessment (12 ranked their familiarity < 6). Those neuropsychologists who provided only Original classifications ($n = 15$) did not differ significantly from those who only provided Reliability classifications ($n = 4$), or those who provided both Original and Reliability classifications ($n = 9$), with respect to their years of experience working as psychologists [$F(2,25) = .87, p < .43$], years of experience with older adult clients [$F(2,25) = 1.49, p < .24$], percentage of caseload made up of older adults [$F(2,25) = .69, p < .51$], or familiarity with the individual measures used in this study [$F(2,25) = .06-1.9, p < .94-.18$].

Procedure and analyses

All information available to the Original neuropsychologist was sent to the second neuropsychologist for reclassification of each case; the second neuropsychologist was blind to the Original classification, but was asked to follow the same classification procedures as the Original neuropsychologist.

The inter-rater agreement for CI classifications between the Original and Reliability neuropsychologists was assessed by calculating a kappa coefficient and percentage agreement.

We identified concordant cases as those given the same CI classification by the Original and Reliability neuropsychologists. The remainder, those where CI classifications differed, were identified as discordant cases. The concordant and discordant cases were compared on measures of participant characteristics including: language of assessment (i.e., English or French) and psychometricians' ratings of participant hearing, vision, cooperativeness, fatigability, inattention, perseveration, impulsivity, social impropriety, tangentiality, physical impairments, and facility with testing using χ^2 tests.

Concordant cases were then recategorized as concordant, not impaired (C-NI) and concordant, impaired (C-I) for comparison with discordant cases. These groups were compared on measures of the participants' cognitive functioning (i.e., 3MS and neuropsychological measures) using one-way analyses of variance (ANOVAs). Even though the neuropsychological test scores were corrected for age, education, and gender, differences might be still expected between the groups for these variables because their influence may not be the same for each group (Reitan & Wolfson, 1996).

Finally, the discordant and concordant groups were compared with respect to rater characteristics using χ^2 tests.

First, to determine whether the confidence rating for the discordant and concordant groups differed, they were compared on the lowest confidence rating provided between the two neuropsychologists rating each case. Subsequently, the groups were compared with respect to the lowest amount of rater experience reported between the two neuropsychologists' ratings for each case, the number of years experience with older adults, and the reported percent of caseload made up of older adults.

RESULTS

Table 1 shows that there was 77% agreement between the Original and Reliability classifications, resulting in a kappa value of .49. When the concordant and discordant cases were compared using χ^2 tests, there were no differences with respect to language in which the assessment was conducted, or any of the ratings made by the psychometricians at the time of testing (see Table 2). When the cases were recategorized as C-NI, C-I, and discordant (D) and compared using ANOVAs, the groups did not differ on age [$F(2,114) = .27, p < .77$]; education [$F(2,114) = 3.57, p < .03$]; CES-D [$F(2,90) = .49, p < .62$]; WRAT-3 [$F(2,96) = .98, p < .38$]; or gender, $\chi^2 = 1.07, p < .59$ (see Table 3). However, the anticipated pattern of performance (i.e., C-NI > discordant > C-I) was seen for the 3MS scores and all 8 of the neuropsychological measures (see Table 4). For the 3MS and 4 of the neuropsychological measures, the C-I group differed significantly ($p < .001$) from both the C-NI and D groups (see Table 4). In all cases, the C-NI and D groups were not distinguishable (see Table 4).

Finally, when the discordant and concordant groups were compared with respect to rater characteristics, there were no differences with respect to the following: lowest confidence rating provided between the two neuropsychologists rating each case [$\chi^2(3, N = 117) = 0.99, p < .80$]; the lowest amount of rater experience reported between the two neuropsychologists' raters (i.e., 5 years or less vs. 6 years or more), [$\chi^2(1, N = 117) = .00, p < .96$]; years of experience with older adults (i.e., 5 years or less vs. 6 years or more), [$\chi^2(1, N = 117) = 0.01, p < .91$]; or percent of caseload made up of older adults [$\chi^2(3, N = 117) = 0.30, p < .59$].

Table 1. Agreement between the original and reliability neuropsychological classifications of no cognitive impairment (NCI) or cognitive impairment (CI)

| | | Reliability classification | | Total |
|-------------------------|-------|----------------------------|----|-------|
| | | NCI | CI | |
| Original classification | NCI | 24 | 11 | 35 |
| | CI | 15 | 67 | 82 |
| | Total | 39 | 78 | 117 |

Table 2. Comparison of concordant and discordant cases based on patient characteristics

| Characteristic | Concordant | Discordant | Chi square | <i>p</i> value |
|--|------------|------------|------------|----------------|
| Language of assessment (<i>N</i> = 117) ^a | 15.4 | 11.5 | .24 | .62 |
| Hearing (<i>N</i> = 116) ^b | 8.9 | 11.5 | .17 | .69 |
| Vision (<i>N</i> = 116) ^b | 10.0 | 7.7 | .13 | .72 |
| Fatigability (<i>N</i> = 117) ^b | 5.5 | 3.8 | .11 | .73 |
| Inattention (<i>N</i> = 117) ^b | 5.5 | 7.7 | .17 | .68 |
| Perseveration (<i>N</i> = 117) ^b | 4.4 | 0 | 1.18 | .28 |
| Impulsivity (<i>N</i> = 117) ^b | 4.4 | 3.8 | .02 | .90 |
| Tangentiality (<i>N</i> = 117) ^b | 3.3 | 0 | .88 | .35 |
| Physical impairments (<i>N</i> = 115) ^c | 8.9 | 12.0 | .22 | .64 |
| Facility with language of testing (<i>N</i> = 117) ^d | | | 3.74 | .29 |
| Cooperativeness (<i>N</i> = 117) ^d | | | 5.52 | .14 |

Note. No participant showed social impropriety.

^a% French.

^b% Problem.

^c% Yes.

^dMeasured on a Likert scale, thus % present or absent is not available.

DISCUSSION

Despite the complexity and the growing importance of identifying the full spectrum of cognitive impairments in older adults, few studies have examined the reliability with which this can be accomplished. Our study is unique in its size (28 neuropsychologists, 117 cases) and scope, in that factors influencing the decision-making of neuropsychologists were examined. In our study, the neuropsychologists showed a

moderate level of agreement when asked to judge whether a participant exhibited CI or not. The level of agreement observed ($\kappa = 0.49$) is virtually identical to that reported in the existing literature for psychologists making similar judgments concerning the presence of cognitive impairment (i.e., 0.48, White et al., 2002) and for dementia diagnoses made by neurologists using clinical records (i.e., 0.49, Solari et al., 1994). When agreement on diagnosis of specific disorders has been examined, agreement tends to be lower (e.g., White et al., 2002). Studies that have reported higher inter-rater agreements have used either well-defined sets of criteria (e.g., DSM-III-R) and standardized written vignettes drawn from carefully selected clinical groups (O'Connor et al., 1996), or a multidisciplinary consensus process (Graham et al., 1996).

Only one clear difference between our concordant and discordant groups was apparent: performance on the cognitive measures. The discordant group fell between the C-NI and C-I groups with respect to cognitive performance on all measures. When significant differences were observed, the discordant group did not differ from the C-NI, whereas the C-I cases were clearly distinguishable from both the C-NI and discordant groups. Of interest, approximately half (12/26) of the discordant cases were persons who received a consensus diagnosis of CIND, a sizable proportion of whom may be in the very early or prodromal stages of a progressive dementia (e.g., Tuokko & Frerichs, 2000; Tuokko et al., 2003). Five of the discordant cases were persons who received a consensus diagnosis of NCI and 9 were lost to follow-up and did not receive a consensus diagnosis. It will be important to determine if the discordant cases are at increased risk for future cognitive decline.

It is possible that the observed agreement rates were reflective of various aspects of the CSHA-3 study design including participant selection criteria and neuropsychological assessment procedures. In the CSHA-3, neuropsychological examinations were conducted for people with CSHA-2

Table 3. Demographic characteristics of groups for whom the neuropsychological classifications were concordant-not impaired (C-NI), discordant (D), or concordant-impaired (C-I)

| Characteristic | <i>N</i> | Mean | <i>SD</i> |
|--------------------|----------|-----------|------------|
| Age | | | |
| C-NI | 24 | 83.9 | 5.2 |
| D | 26 | 84.4 | 6.2 |
| C-I | 67 | 84.9 | 5.7 |
| Years of education | | | |
| C-NI | 24 | 8.7 | 3.1 |
| D | 26 | 8.5 | 3.1 |
| C-I | 67 | 10.3 | 3.6 |
| CES-D | | | |
| C-NI | 22 | 8.3 | 2.8 |
| D | 22 | 8.7 | 4.5 |
| C-I | 49 | 9.2 | 3.9 |
| WRAT-3 | | | |
| C-NI | 17 | 47.1 | 5.2 |
| D | 22 | 43.8 | 8.8 |
| C-I | 60 | 44.7 | 7.8 |
| Gender | | Male (%) | Female (%) |
| C-NI | 24 | 11 (45.8) | 13 (54.2) |
| D | 26 | 9 (34.6) | 17 (65.4) |
| C-I | 67 | 23 (34.3) | 44 (65.7) |

Note. CES-D = The Center for Epidemiological Studies-Depression Scale; WRAT-3 = Wide Range Achievement Test, Reading Subtest.

Table 4. Means and standard deviations (SD) for measures of cognitive performance between groups for whom the neuropsychological classifications were concordant–not impaired (C-NI), discordant (D), or concordant–impaired (C-I)

| Measure | <i>N</i> | Mean | <i>SD</i> | <i>F</i> | <i>p</i> |
|-------------------------------|----------|-------|-----------|----------|----------|
| 3MS | | | | | |
| C-NI | 24 | 86.17 | 4.67 | 12.31 | .000 |
| D | 26 | 84.77 | 7.02 | | |
| C-I* # | 67 | 78.54 | 8.44 | | |
| WMS-I | | | | | |
| C-NI | 24 | 4.87 | 0.74 | 14.21 | .000 |
| D | 26 | 4.77 | 1.07 | | |
| C-I* # | 67 | 4.08 | 1.52 | | |
| BCR-T1, T-score | | | | | |
| C-NI | 13 | 53.47 | 10.26 | 22.05 | .000 |
| D | 21 | 50.30 | 7.82 | | |
| C-I* # | 51 | 30.99 | 16.51 | | |
| BCR-T, T-Score | | | | | |
| C-NI | 13 | 53.47 | 10.26 | 22.05 | .000 |
| D | 21 | 50.30 | 7.82 | | |
| C-I* # | 51 | 30.99 | 16.51 | | |
| WAISR-S, T-score | | | | | |
| C-NI | 13 | 54.46 | 10.67 | 3.61 | .032 |
| D | 21 | 47.80 | 10.28 | | |
| C-I | 51 | 45.54 | 10.87 | | |
| FAS, T-score | | | | | |
| C-NI | 13 | 50.80 | 11.81 | 4.67 | .012 |
| D | 21 | 46.15 | 9.22 | | |
| C-I | 51 | 41.06 | 11.48 | | |
| Animal naming, T-score | | | | | |
| C-NI | 13 | 53.09 | 15.22 | 10.95 | .000 |
| D | 21 | 46.05 | 9.23 | | |
| C-I* | 51 | 38.57 | 9.90 | | |
| WAISR-BD, T-score | | | | | |
| C-NI | 13 | 50.02 | 6.84 | 5.43 | .006 |
| D | 21 | 47.19 | 8.79 | | |
| C-I | 51 | 41.94 | 9.54 | | |
| WAISR-DS, T-score | | | | | |
| C-NI | 13 | 53.47 | 11.17 | 8.83 | .000 |
| D | 20 | 53.18 | 12.55 | | |
| C-I* # | 48 | 41.17 | 13.21 | | |
| BNT, Total score | | | | | |
| C-NI | 23 | 23.57 | 4.67 | 4.37 | .015 |
| D | 25 | 20.48 | 5.37 | | |
| C-I | 65 | 16.69 | 5.65 | | |

Note. * indicates significant ($p < .05$) difference from C-NI; # indicates difference from D; 3MS = Modified Mini-Mental State Examination; WMS-I = Wechsler Memory Scale–Information; BCR-T1 = Buschke Cued Recall, Free Recall Trial 1; BCR-T = Buschke Cued Recall, Total Free Recall; WAISR-S = WAISR Similarities; FAS = Controlled Word Association Test; WAISR-BD = WAISR Block Design; WAISR-DS = WAISR Digit Symbol; BNT = Boston Naming Test.

clinical diagnoses of NCI or CIND and for all participants with CSHA-3 3MS scores falling above 49 but below 90. These restrictions, placed by the 3MS selection criteria, truncated the range of the participants' cognitive functioning at the upper and lower ends where concordance rates may be expected to be highest, thereby reducing the overall concordance rates.

Another aspect of the study design that may have influenced agreement rates was the limited information about the participant available to the neuropsychologists. For example, under more ideal circumstances, neuropsychologists might have access to medical information about the person at the time that classifications are being made. In our situation, only self-report of health conditions and health status was available. It is possible that access to medical information would have increased agreement. However, White et al. (2002) had access to medical information and obtained a virtually identical rate of agreement to ours.

Despite these limitations, our finding that disagreements between the neuropsychologists occur primarily for cases with subtle, or mild, CI reinforces the need for additional research in this area. Much controversy currently exists regarding the nosology of such cognitive impairment. Some researchers promote a restrictive definition of mild CI, characterized by memory impairment in the context of normal general cognitive function (e.g., Maruff et al., 2004; Petersen et al., 1999; 2001). Other researchers, principally those working in Europe (e.g., Ritchie & Touchon, 2000; Ritchie et al., 2001; Touchon & Ritchie, 1999), favor a more inclusive nosological entity such as Age-Associated Cognitive Decline (AACD; Levy, 1994) that refers to a decline in any cognitive functions (attention, learning and memory, thinking, language, and visuospatial function), and is identified in relation to norms for elderly persons. More recently, Petersen (2003, 2004) has proposed various subtypes of mild CI including amnesic, single domain nonmemory, and multiple domain. Not until neuropsychologists come to terms with the classification of mild forms of CI and articulate the best ways to define them, will inter-rater agreement improve beyond our present moderate rates.

ACKNOWLEDGMENTS

A research personnel award from the Canadian Institutes of Health Research, Institute of Aging supported HT in the preparation of this manuscript. Phases 1 and 2 of the Canadian Study of Health and Aging core study were funded by the Seniors' Independence Research Program, through Health Canada's National Health Research and Development Program [NHRDP project #6606-3954-MC(S)]; supplementary funding for analysis of the caregiver component was provided by the Medical Research Council. Additional funding was provided by Pfizer Canada Incorporated, through the Medical Research Council/Pharmaceutical Manufacturers Association of Canada Health Activity Program, NHRDP [project #6603-1417-302(R)], by Bayer Incorporated, and by the British Columbia Health Research Foundation [projects #38 (93-2) and #34 (96-1)]. Core funding for Phase 3 was obtained from the Canadian Institutes for Health Research (CIHR grant # MOP-42530); supplementary funding for the caregiver component was obtained from CIHR grant # MOP-43945. Additional funding was provided by Merck-Frosst and by Janssen-Ortho Inc. The study was coordinated through the University of Ottawa and Health Canada. We would specifically like to acknowledge the assistance of Liz Sykes, Nansy Jean-Baptiste, Maggie Stewart, and Beth Sander, who worked very hard to prepare the materials for the reliability review, selected the sample, prepared the clinical examination summaries,

worked with the volunteer neuropsychologists, and prepared the data for analysis.

REFERENCES

- Andresen, E.M., Malmgren, J.A., Carter, W.B., & Patrick, D.L. (1994). Screening for depression in well older adults: Evaluation of a short form of the CES-D. *American Journal of Preventive Medicine, 10*, 77–84.
- Baldereschi, M., Amato, M.P., Nencini, P., Pracucci, G., Lippi, A., Amaducci, L., Gauthier, S., Beatty, L., Quiroga, P., & Klassen, G. (1994). Cross-national interrater agreement on the clinical diagnostic criteria for dementia. WHO-PRA Age-associated Dementia Working group, WHO-program for Research on Aging, Health of Elderly Program. *Neurology, 44*, 239–242.
- Ballantyne, B.T., French, A.K., Heimsoth, S.L., Kachingwe, A.F., Lee, J.B., & Soderberg, G.L. (1995). Influence of examiner experience and gender on interrater reliability of KT-10000 arthrometer measurements. *Physical Therapy, 75*, 898–906.
- Brooks, D. & Thomas, J. (1995). Interrater reliability of auscultation of breath sounds among physical therapists. *Physical Therapy, 75*, 1082–1088.
- Canadian Study of Health and Aging Working Group. (1994). Canadian Study of Health and Aging: Study methods and prevalence of dementia. *Canadian Medical Association Journal, 150*, 899–913.
- Fisher, N.J., Tierney, M.C., & Snow, W.G. (1999). Odd/even short forms of the Boston Naming Test: Preliminary geriatric norms. *Clinical Neuropsychologist, 13*, 359–364.
- Graham, J.E., Rockwood, K., Beattie, B.L., McDowell, I., Eastwood, R., & Gauthier, S. (1996). Standardization of the diagnosis of dementia in the Canadian Study of Health and Aging. *Neuroepidemiology, 15*, 246–256.
- Graham, J.E., Rockwood, K., Beattie, B.L., Eastwood, R., Gauthier, S., Tuokko, H., & McDowell, I. (1997). Prevalence and severity of cognitive impairment with and without dementia in an elderly population. *Lancet, 349*, 1793–1796.
- Hogervorst, E., Barnetson, L., Jobst, K.A., Nagy, Z., Combrinck, M., & Smith, A.D. (2000). Diagnosing dementia: Interrater reliability assessment and accuracy of the NINCDS-ADRDA criteria versus CERAD histopathological criteria for Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders, 11*, 107–113.
- Jorm, A.F., Rodgers, B., Henderson, A.S., Korten, A.E., Jacomb, P.A., Christensen, H., & Mackinnon, A. (1998). Occupation type as a predictor of cognitive decline and dementia in old age. *Age and Ageing, 27*, 477–483.
- Kukull, W.A., Larson, E.B., Reifler, B.V., Lampe, T.H., Yerby, M., & Hughes, J. (1990). Interrater reliability of Alzheimer's disease diagnosis. *Neurology, 40*, 257–260.
- Levy, R. (1994). Age-associated cognitive decline. Working Party of the International Psychogeriatric Association in collaboration with the World Health Organization. *International Psychogeriatrics, 6*, 63–68.
- Lopez, O.L., Swihart, A.A., Becker, J.T., Reinmuth, O.M., Reynolds, C.F., Rezek, D.L., & Daly, F.L. (1990). Reliability of the NINCDS-ADRDA clinical criteria for the diagnosis of Alzheimer's disease. *Neurology, 40*, 1517–1522.
- Maruff, P., Collie, A., Darby, D., Weaver-Cargin, J., & Masters, C. (2004). Subtle memory decline over 12 months in mild cognitive impairment. *Dementia and Geriatric Cognitive Disorders, 18*, 342–348.
- O'Connor, D.W., Blessed, G., Cooper, B., Jonker, C., Morris, J.C., Presnell, I.B., Ames, D., Kay, D.W.K., Bickel, H., Schäufele, M., Wind, A., Coats, M., & Berg, L. (1996). Cross-national interrater reliability of dementia diagnosis in the elderly and factors associated with disagreement. *Neurology, 47*, 1194–1199.
- O'Connor, D.W., Pollitt, P.A., & Treasure, F.P. (1991). The influence of education and social class on the diagnosis of dementia in a community population. *Psychological Medicine, 21*, 219–224.
- Petersen, R.C. (2003). *Mild cognitive impairment: Aging to Alzheimer's disease*. New York: Oxford University Press.
- Petersen, R.C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine, 256*, 183–194.
- Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., & Kokmen, E. (1999). Mild cognitive impairment. *Archives of Neurology, 56*, 303–308.
- Petersen, R.C., Stevens, J.C., Ganguli, M., Tangalos, E.G., Cummings, J.L., & DeKosky, S.T. (2001). Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology, 56*, 1133–1142.
- Reitan, R.M. & Wolfson, D. (1996). Relationships of age and education to Wechsler Adult Intelligence Scale IQ values in brain-damaged and non-brain damaged groups. *Clinical Neuropsychologist, 10*, 293–304.
- Ritchie, K., Artero, S., & Touchon, J. (2001). Cognitive classification criteria for mild cognitive impairment: A population-based validation study. *Neurology, 56*, 37–42.
- Ritchie, K. & Touchon, J. (2000). Mild cognitive impairment: Conceptual basis and current nosological status. *Lancet, 355*, 225–228.
- Rosen, W.G. (1980). Verbal fluency in aging and dementia. *Journal of Clinical Neuropsychology, 2*, 135–146.
- Satz, P. & Mogel, S. (1962). An abbreviation of the WAIS for clinical use. *Journal of Clinical Psychology, 18*, 77–79.
- Shapiro, E. & Tate, R.B. (1997). The use and cost of community care services by elders with unimpaired cognitive function, with cognitive impairment/no dementia and with dementia. *Canadian Journal on Aging, 16*, 665–681.
- Solari, A., Cossa, F.M., Denes, F., Gainotti, G., Grossi, D., Nichelli, P., & Filippini, G. (1994). Agreement in clinical diagnosis of dementia: Evaluation of a case series with mild cognitive impairment. *Neuroepidemiology, 13*, 89–96.
- Spree, O. & Benton, A.L. (1977). *Neurosensory Center Comprehensive Examination for Aphasia: Manual of instructions*. Victoria, B.C.: University of Victoria.
- Steinhausen, H.C. & Erdin, A. (1991). The inter-rater reliability of child and adolescent psychiatric disorders in the ICD-10. *Journal of Child Psychology and Psychiatry, 32*, 921–928.
- Stern, Y., Gurland, B., Tatemichi, T.K., Tang, M.X., Wilder, D., & Mayeux, R. (1994). Influence of education and occupation on the incidence of Alzheimer's disease. *Journal of the American Medical Association, 271*, 1004–1010.
- Stones, M.J., Kozma, A., Hirdes, J., Gold, D., Arbuckle, T., & Kolopack, P. (1996). Short Happiness and Affect Research Protocol (SHARP). *Social Indicators Research, 37*, 75–91.
- Teng, E.L. & Chui, H.C. (1987). The Modified Mini-Mental State (3MS) Examination. *Journal of Clinical Psychiatry, 48*, 314–318.
- Touchon, J. & Ritchie, K. (1999). Prodromal cognitive disorder in

- Alzheimer's disease. *International Journal of Geriatric Psychiatry*, *14*, 556–563.
- Tuokko, H. & Frerichs, R. (2000). Cognitive impairment with no dementia (CIND): Longitudinal studies, the findings, and the issues. *The Clinical Neuropsychologist*, *14*, 504–525.
- Tuokko, H., Frerichs, R., Graham, J., Rockwood, K., Kristjansson, B., Fisk, J., Bergman, H., Kozma, A., & McDowell, I. (2003). Five-year follow-up of cognitive impairment with no dementia. *Archives of Neurology*, *60*, 577–582.
- Tuokko, H., Kristjansson, E., & Miller, J.A. (1995). Neuropsychological detection of dementia: An overview of the neuropsychological component of the Canadian Study of Health and Aging. *Journal of Clinical and Experimental Neuropsychology*, *17*, 353–373.
- Tuokko, H., Vernon-Wilkinson, R., Weir, J., & Beattie, B.L. (1991). Cued recall and early identification of dementia. *Journal of Clinical and Experimental Neuropsychology*, *13*, 871–879.
- Tuokko, H. & Woodward, T.S. (1996). Development and validation of a demographic correction system for neuropsychological measures used in the Canadian Study of Health and Aging. *Journal of Clinical and Experimental Neuropsychology*, *18*, 479–616.
- Wechsler, D. (1975). *Wechsler Memory Scale*. New York: Psychological Corporation.
- Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised*. New York: Psychological Corporation.
- White, R.F., James, K.E., Vasterling, J.J., Marans, K., Delaney, R., Kregel, M., & Rose, F. (2002). Interrater reliability of neuropsychological diagnoses: A Department of Veterans Affairs cooperative study. *Journal of the International Neuropsychological Society*, *8*, 555–565.
- Wilkinson, G.S. (1993). *Wide Range Achievement Test (WRAT-3) administration manual*. Wilmington, DE: Wide Range, Inc.