The neuropsychometrics of aging

Normative studies in the Maastricht Aging Study

Neuropsych Publishers, Maastricht, The Netherlands

Printed by Datawyse BV, Maastricht

ISBN-10: 90-75579-29-2

ISBN-13: 978-90-75579-29-1

© W. Van der Elst, Maastricht 2006

All rights are preserved, whether the whole or part is concerned. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the copyright owner.

Neuropsych Publishers is a non-profit organization, which aims at promoting the science of 'Brain and Behavior' and improving the application of the products of this science in health care and education. Neuropsych Publishers accomplishes these aims by publishing books, dissertations and other products of scientific activity, by disseminating educational material and publication of tests, assessment scales and other psychometric instruments in the field of Neuropsychology, Neuropsychiatry and other areas within the domain of Brain and Behavior.

Postal address: Neuropsych Publishers, Department of Psychiatry and Neuropsychology, Maastricht University, P.O. Box 616, NL-6200 MD Maastricht, The Netherlands, www-np.unimaas.nl

The neuropsychometrics of aging

Normative studies in the Maastricht Aging Study

PROEFSCHRIFT

Ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van Rector Magnificus,
Prof. mr. G. P. M. F. Mols
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op donderdag 12 oktober 2006 om 12:00 uur

door

Wim Van der Elst

Promotor

Prof. Dr. J. Jolles

Copromotor

Dr. M. Van Boxtel

Beoordelingscommissie

Prof. Dr. M. Prins (voorzitter)

Prof. Dr. E. de Haan (Universiteit Utrecht)

Dr. B. Jansma

Prof. Dr. H. Merckelbach

Prof. Dr. B. Schmand (Vrije Universiteit Amsterdam)

• Mondriaan Zorggroep • Vijverdal • RIAGG Maastricht
• RGC Westelijke Mijnstreek • RIBW Heuvelland en Maasvallei
• Academisch Ziekenhuis Maastricht • Atrium Medisch Centrum
• Universiteit Maastricht • Onderzoeksschool Experimentele Psychopathologie
• Onderzoeksschool Hersenen en Gedrag

• de Wetenschapskring bevordert wetenschappelijk onderzoek op het gebied van de GGZ in Zuid-Limburg •

The research described in this thesis was performed at the Maastricht Brain and Behavior Institute, Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, The Netherlands.

Financial support for the publication of this thesis has been kindly provided by **Alzheimer Nederland, Pfizer bv, Wyeth Pharmaceuticals bv and Harcourt Test Publishers.**

Contents

<i>Chapter 1</i> Introduction	7
<i>Chapter 2</i> The Stroop Color-Word Test: influence of age, sex, and education; and normative data for a large sample across the adult age range	15
<i>Chapter 3</i> Rey's Verbal Learning Test: normative data for 1,855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation	39
<i>Chapter 4</i> Normative data for the animal, profession and Letter m naming Verbal Fluency Tests for Dutch speaking participants and the effects of age, education, and sex	59
<i>Chapter 5</i> The Concept Shifting Test: adult normative data	73
<i>Chapter 6</i> Assessment of information processing in working memory in applied settings: the Paper & Pencil Memory Scanning Test	89
<i>Chapter 7</i> The Letter Digit Substitution Test: normative data for 1,858 healthy participants aged 24-81 from the Maastricht Aging Study (MAAS): influence of age, education and sex	107
<i>Chapter 8</i> Assessment of change in performance on the Stroop Color-Word Test, Verbal Learning Test of Rey, and Letter Digit Substitution Test after a test-retest interval of three years	117
<i>Chapter 9</i> Concluding Remarks	135
<i>References</i>	143
<i>Summary</i>	151
<i>Samenvatting</i>	155
<i>Curriculum vitae/Publications</i>	159

Introduction

At present, about 15% of the European population is older than 65, and this proportion is expected to rise sharply to about one in four people in 2050 (Beets, 2000). As the number of older people in Western societies increases, so does the number of people who suffer from pathological cognitive disorders such as dementia (the incidence of dementia increases exponentially with age, with no levelling off in very old age; Jorm & Jolley, 1998). Due to the increased numbers of older and demented people in western societies and consequently the increasing numbers of people who are referred for cognitive evaluation in the clinic (LaRue, 1992), the cognitive and psychological aspects of aging became increasingly more studied in the last few decades with an exponential growth of publications on aging-research.

Neuropsychological assessment is a method to examine the brain by studying its behavioural product (Lezak et al., 2004; pp. 15). Neuropsychological assessment is widely used in clinical settings and in theoretical research, for example when evaluating cognitive impairment in older people suspected of conditions such as dementia (Petersen et al., 2001), or when conducting research that involves the measurement and explanation of the mechanisms, patterns, onset, and magnitude of age-related changes in cognition (Hertzog, 1996). There is a wide variety of neuropsychological instruments available (for an overview, see Lezak et al., 2004 or Spreen & Strauss, 1998), but despite the frequent use of these instruments in both clinical and research settings, there is a paucity of systematic studies into the psychometric characteristics of these tests. Yet, there is a great need for this type of studies, especially because norms for important tests and cognitive domains are not available. As a consequence, appropriate interpretation of neuropsychological test results is hindered and misinterpretation of the cognitive status of an individual may occur.

An understanding of what constitutes “normal” performance is needed before an opinion regarding a person’s cognitive strengths and weaknesses of various neurobehavioral capacities can be given (Mitrushina et al., 1999; pp. 3). Establishing the normal range of performance is not only relevant from a psychometric viewpoint, but also provides insight into theoretical questions that are often studied in cognitive gerontology, such as questions with respect to the factors that are related to inter-individual differences in cognitive functioning, and questions regarding the factors that affect changes in cognitive functioning over time. As the samples that are used in psychometric studies are typically large and representative, the observed relationships between demographical variables and performance on cognitive tests can be considered to be robust (Salthouse, 1996).

In the next sections, the process of cognitive aging is described, some background information on neuropsychological assessment and neuro-psychometrics

is provided, the contribution of normative studies to theoretical questions in gerontology is exemplified, some background information regarding the Maastricht Aging Study is given, and the problem statement of this thesis is clarified. Finally, the aims and approach are described, and an outline of the studies described in this thesis is given.

Cognitive aging

Most studies that evaluated age-related differences in cognition used cross-sectional designs (Birren & Schroots, 1996). In cross-sectional studies, people of different ages are compared at the same point in time. The modal implementation of a cross-sectional design is the use of extreme-age-groups, in which the average performance of a sample of young adults is compared with the average performance of a group of older adults (Herzog, 1996). Other implementations extend this design to full cross-sectional studies across the entire adult age span (e.g. the studies by Houx et al., 1991, who used healthy people aged 20, 30, 40 up till 80 years of age). The bulk of these studies consistently showed that cross-sectional age differences are not unitary across cognitive domains. A fruitful distinction of cognitive domains into fluid and crystallized cognitive abilities was made by Cattell (1963; see also Horn, 1982). Fluid abilities reflect maturational growth and decline of neuronal structures, whilst crystallized abilities depend upon the accumulation of formal and informal educational experiences (LaRue, 1992; pp. 48). Fluid cognitive abilities such as attention, memory, visuo-spatial processing, abstract reasoning, executive functioning, and information processing speed were consistently found to be negatively affected in older people as compared to their younger counterparts (Craik & Salthouse, 2000; LaRue, 1992; Spreen & Strauss, 1998). On the other hand, knowledge-based crystallized activities such as vocabulary knowledge tend to increase up to the sixth or seventh decade (Anstey, 1993; Christensen, 2001). The most authoritative theory to explain the age-related differences in fluid cognitive abilities is the processing speed theory. The central hypothesis of this theory is that the speed with which many processing operations can be executed decreases with increasing age (Salthouse, 1996). This reduction in the speed of information processing is hypothesized to be a major contributor to adult age differences in many measures of cognition due to two mechanisms. The limited time mechanism is assumed to operate because relevant cognitive operations are too slowly to be successfully completed in the available time; the simultaneity mechanism is hypothesized to operate because slow processing reduces the amount of simultaneously available information needed for higher level processing (Salthouse, 1996; pp. 404).

As noted, the bulk of studies that evaluated age-related differences in cognition were cross-sectional. A well-known problem with cross-sectional studies is that age-effects cannot be distinguished from cohort effects (Botwinick, 1977;

Williams & Klug, 1996). For example, age is not the only factor that distinguishes younger from older groups: both groups also differ with respect to other factors that may affect cognition such as the adequacy of nutrition during childhood or educational opportunities. As a consequence, cross-sectional studies tend to overestimate the true effect of age on cognition (Schaie, 1988). More recently, longitudinal studies became increasingly more employed to study age-related changes in cognition (e.g. the Longitudinal Aging Study Amsterdam (Deeg et al., 1993), the Maastricht Aging Study (Jolles et al., 1995), and the Canberra Longitudinal Study of Aging (Logie et al., 2004)), but they are still relatively scarce. Longitudinal studies follow a cohort of people over time with repeated measurements of the variables of interest. Such studies do not confound cohort and age-effects, but these designs have other problems. For example, longitudinal studies tend to underestimate the true age-related change in cognition due to the occurrence of practice effects and selective drop-out (Hay-slip & Panek, 1989). Several longitudinal studies have indeed shown that the effects of aging on cognition were fewer and less pronounced when compared with cross-sectional studies (see e.g. Hulstsch et al., 1998; Schaie, 1996), but the general picture that emerges from longitudinal studies is similar to that obtained in cross-sectional studies. Thus, aging is indeed associated with a decline in fluid cognitive abilities.

Neuropsychological assessment and neuro-psychometrics

A neuropsychological examination provides a comprehensive evaluation of the cognitive domains that are putatively associated with various brain substrates (Mitrushina et al., 1999; pp. 3). Cognitive abilities are never directly observed, but must always be inferred from proxy measures. For example, a person's "verbal memory" cannot be observed directly; what can be observed is the person's ability to recall verbal material in a specific standardized test-setting. Thus, neuropsychological testing consists of using samples of an individual's behaviour to make inferences about his or her cognitive functioning. Since the subject of neuropsychological assessment is behaviour (i.e. performance on a particular test or task), it relies on many of the same assumptions and techniques used in psychological assessment (Lezak et al., 2004).

Theories and techniques of psychological measurement are studied in the field of psychometrics. The theories of measurement in psychometrics can be roughly divided into the Classical Test Theory (CTT; e.g. Gulliksen, 1950) and the Item Response Theory (IRT; e.g. Lord and Novick, 1968). Both approaches differ in various aspects, although many principles of CTT can be derived from IRT (Lord and Novick, 1968). In view of the subject of the present thesis, a particularly interesting difference between both techniques is how score meaning is determined. In the CTT, the test scores of an individual subject only have meaning when they are compared with the test scores of other people for their

relative position. In other words, score meaning is determined by reference to the test scores of a normative sample. In IRT models on the other hand, persons (trait scores) and items (item difficulty) are calibrated on a common scale. Thus, score meaning is referenced directly to the items of the test in the IRT. The comparison between a person's trait score and an item's difficulty has direct meaning for evaluating performance: if a person's trait level exceeds the item difficulty, the person is more likely to pass the item (and vice versa). The use of IRT has been very limited in neuropsychology until now (Mungas & Reed, 2000), but the potential advantages of using IRT instead of CTT should not be overlooked (see Chapter 9).

The contribution of normative studies to theoretical questions in cognitive gerontology

As noted, the CTT is the dominant theory of measurement in neuropsychological assessment. In the CTT, a "raw" test score on a (neuro)psychological test cannot be interpreted in a meaningful way without an empirical frame of reference (Capitani, 1997), i.e. a normative group. Different studies have shown that independent variables such as age, gender, and level of education affect performance on cognitive tasks (e.g. Mitrushina et al., 1999). For example, the 'normal' performance of a 80-year-old low educated male on a verbal memory test is quite different from the normal performance of a 25-year-old highly educated women on the same test (Van der Elst et al., 2005). Thus, the first step in normative studies is to evaluate to what extent test performance is affected by factors such as age, gender, and education. Evaluating the influence of demographical variables on performance on cognitive tests is not only relevant in the context of normative analyses, but also provides insight in theoretical issues in cognitive gerontology. For example, a topic of interest in cognitive gerontology is whether there are individual differences in cognitive aging trajectories. By formal modelling of the effects of age, gender, and education on neuropsychological test scores – as is done in normative research – the relevance and relative contribution of each of these factors on change in cognitive abilities can be evaluated.

The Maastricht Aging Study

The Maastricht Aging Study (MAAS; Jolles et al., 1996) is a longitudinal program that was primarily devised to find answers regarding questions about the normal – i.e. non pathological – aging of cognitive functions.

The baseline measurement in MAAS involved almost 2,000 cognitively intact and healthy people over the entire adult age range (24 - 81 years). Baseline measurements were conducted between 1993 and 1996, with four follow-up

waves being scheduled. The follow-up frequency in MAAS was dependent on the age at first measurement, i.e. 3 years for people aged 49 years and older, and 6 years for people younger than 49. Major dependent variables in MAAS are cognitive, making MAAS an essentially cognitive gerontological program. In addition to cognitive performance, MAAS yields a wealth of answers on the domain of medical and sensory functioning (e.g. van Boxtel et al., 1997; van Boxtel et al., 2000; van Hooren et al., 2005; Valentijn et al., 2005), psychosocial functioning (e.g. Bosma et al., 2002) and psychological factors (e.g. Ponds et al., 2000; Mol et al., 2006). Three main research questions of MAAS on the domain of cognitive gerontology are: When in the course of a human lifetime do changes in memory and memory-related functions appear? Do all cognitive functions deteriorate at the same rate or are specific aspects of cognition affected more than others? And, do individuals differ with respect to the age of onset and the rate of progression on cognitive aging? (Jolles et al., 1996; pp. 17-19). Assuming that some participants, who were healthy and cognitively intact at baseline, will exhibit abnormalities at follow-up measurements, the longitudinal part of MAAS allows for studying pathological forms of aging and searching for prodromes of dementia and other classes of psychopathology such as depression.

Problem statement

Normative data provide an external criterion to evaluate the raw scores on neuropsychological tests and to make inferences with respect to the integrity of the unobserved, or latent, underlying cognitive abilities. Norms are a fundamental tool in neuropsychological assessment: clinicians or researchers refer to norms whenever they wonder whether a deficit is present in a cognitive function (Capitani, 1997). Although clinicians and researchers acknowledge the importance of normative data, there are only few well-normed (neuro)psychological tests – i.e. tests with norms that are based on a large representative sample of healthy individuals with appropriate corrections for demographic influences. Normative data that evaluate change in cognitive performance in individuals are almost completely absent. The following factors are, at least in part, responsible for the lack of appropriate norms for neuropsychological tests.

First, many neuropsychological tests are poorly standardized. For example, there are about 20 different administration procedures of the Auditory Verbal Learning Test of Rey (Schmidt, 1996). Different studies provided normative data for the different administration variants, but these norms are usually based on small samples. Due to the lack of standardization, it is problematic to combine the data of the different studies into a large meta-sample to obtain reliable norms.

Secondly, the traditional statistical approach of establishing normative data is not optimal (Van Breukelen & Vlaeyen, 2005; Van der Elst et al., 2006a). In

the traditional approach, summary statistics such as the means and standard deviations of the raw scores on a test are computed. These statistics are used to convert a person's raw score into an easily interpretable measure such as a Z-score. One important drawback of the traditional approach is that the total sample usually needs to be broken down into various subgroups, because the performance on many neuropsychological tests is influenced by factors such as age, sex, and education, amongst others (e.g. Lezak et al., 2004; Mitrushina et al., 1999). As a result, the estimated population means and population variances for the different subgroups are based on small sub-samples (n) of the total normative sample (N), thus giving rise to unreliable norms due to chance trends in the data.

Thirdly, the test scores of various neuropsychological tests are affected by culture and/or language (Ardila, 1995). As a consequence, normative data that were collected with English-speaking people are not optimal for use in other languages. Yet, quite some norms which have been obtained in studies of white middleclass North American subjects are used also in Spanish speaking subjects in the US or in people in other countries such as Dutch speaking people in the Netherlands.

Aims

The aim of the present thesis is twofold; the first aim is to evaluate the influence of age and age-extrinsic factors on inter-individual differences in the performance on cognitive tests (using cross-sectional data), as well as the influence of these variables on changes in performance on cognitive tests over time (using longitudinal data). The second aim is to establish the normal range of performance for three commonly used neuropsychological tests (i.e. the Stroop Colour-Word Tests, the Verbal Learning Test of Rey, and the Verbal Fluency Test) and for three newly developed tasks (the Concept Shifting Tests, the Paper & Pencil Memory Scanning Test, and the Letter Digit Substitution Test). These tests quantify cognitive abilities that are typically assessed in most neuropsychological examinations, i.e. short-term verbal memory and learning, long-term verbal memory, recognition memory, speed of information processing, inhibition, executive functions, concept shifting, verbal fluency and working memory.

Approach

The principles of the CTT rather than the principles of IRT are used in the present thesis to comply with the bulk of the literature on neuropsychological assessment and to allow for a comparison of our results with previous research. In this approach, score meaning is determined by reference to a large sample of cognitively intact people. The MAAS study provides an excellent sample to

conduct normative analyses because it includes a large number of cognitively intact males and females of various educational levels over the full adult age range. We used a regression-based normative approach to evaluate the influence of demographical variables on inter-individual differences in the performance on cognitive tests as well as the influence of these variables on changes in performance on cognitive tests over time. Data from the baseline measurement of MAAS were used to establish regular normative data (i.e. norms for a single measurement), MAAS follow-up data were used to establish the norms for change (i.e. norms for the change between two repeated measures).

Outline of the thesis

Each of the Chapters 2 – 7 evaluated the influence of age and age-extrinsic variables on differences in performance on specific neuropsychological tests (the Stroop Color-Word Test, the Verbal Learning Test, the Verbal Fluency Test, the Concept Shifting Test, the Memory Scanning Test and the Letter Digit Substitution Test). For each test, the normal range of test-performance was established. Chapter 8 evaluated the change in performance with a test-retest interval of three years in a sample of people aged above 50, and provided normative change data.

Chapter 2 evaluated the effects of demographical variables on the Stroop Color-Word Test (SCWT; Stroop, 1935) and established the normal range of performance. The SCWT measures executive functioning (Moering et al., 2004) and related concepts such as attention and cognitive flexibility. Both a traditional and a regression-based normative approach was used, and the appropriateness of both methods was evaluated.

Chapter 3 evaluated the effects of demographical variables and mode of presentation (visual versus auditory presentation of the stimuli) on the normal range of performance for the Verbal Learning Test (VLT; Rey, 1958). The VLT assesses short-term memory, learning capacity, long-term memory retrieval of newly learned verbal material, and recognition memory (Lezak et al., 2004).

Chapter 4 provided normative data for semantic and phonemic Verbal Fluency Tests (VFTs). Previous research showed that performance on VFTs is strongly influenced by culture and language, so it was necessary to establish norms for Dutch speaking people. The VFT is mostly used to measure executive functioning (Henry & Crawford, 2004; Mitrushina et al., 1999), but VFTs have also been used to measure other functions such as word knowledge, access to semantic memory, long-term verbal memory, and attention (Crowe, 1998; Ruff et al., 1997).

Chapter 5 established the normal range of performance for a newly developed test, the Concept Shifting Test (CST). This test is a modification of the more widely known Trail Making Test. The CST was especially developed to avoid certain problems with the original TMT that hamper specificity of the

executive function measure. This test measures concept shifting, which is usually considered as being part of a broader cognitive function, i.e. executive function.

In Chapter 6, a paper-and-pencil version of Sternberg's Memory Scanning Test (P&P MST) was presented. This new test was especially developed for use in the clinic, because the original Sternberg task is difficult to use in a clinical environment. The P&P MST is used to measure memory scanning and non-memory functions in working memory.

Chapter 7 presents normative data for the Letter Digit Substitution Test. This test is an adaptation of the Digit Symbol Substitution Test (Wechsler, 1955, 1981), but has the advantage that the responses of the testee can be given in writing and verbally, and that over-learned symbols are used. The LDST measures general speed of information processing.

In Chapter 8, the influence of demographical variables on the change in performance after three years in the VLT, SCWT and LDST is evaluated, and normative change data are provided.

Chapter 9 provides some concluding remarks.

Chapter 2

The Stroop Color-Word Test: influence of age, sex, and education; and normative data for a large sample across the adult age range

Wim Van der Elst, Martin P. J. Van Boxtel, Gerard J. P. Van Breukelen & Jelle Jolles

Abstract

The Stroop Color-Word Test was administered to 1,856 cognitively screened, healthy Dutch-speaking participants aged 24 to 81 years. The effects of age, sex, and education on Stroop test performance were investigated in order to adequately stratify the normative data. The results showed that especially the speed-dependent Stroop scores (time to complete a subtest), rather than the accuracy measures (the errors made per Stroop subtask), were profoundly affected by the demographic variables. In addition to the main effects of the demographic variables, an Age x low Level of Education interaction was found for the Error III and the Stroop Interference scores. This suggests that executive function, as measured by the Stroop test, declines with age and that the decline is more pronounced in people with a low level of education. This is consistent with the reserve hypothesis of brain aging (i.e., that education generates reserve capacity against the damaging effects of aging on brain functions). Normative Stroop data were established using both a regression-based and traditional approach, and the appropriateness of both methods for generating normative data is discussed.

Introduction

The Stroop Color-Word Test (Stroop, 1935) is a useful and reliable assessment tool in psychology (Lezak, Howieson, & Loring, 2004). Numerous different Stroop test versions have been developed (e.g. Comalli, Wapner, & Werner, 1962; Golden, 1978; Trenerry, Crosson, DeBoe, & Leber, 1989), with variations in the color and number of the test items, the number of subtests, and the administration procedure. Despite these variations, the basic paradigm of the Stroop test has remained the same: an individual's performance on a basic task (for example, reading names of colors) is compared with his/her performance on an analogous task in which a habitual response needs to be suppressed in support of an unusual one (i.e. naming the ink color that incongruously named color words are printed in). The increase in time taken to perform the latter task compared with the basic task is referred to as 'the Stroop interference effect' (e.g. Davidson, Zacks, & Williams, 2003; Moering, Schinka, Mortimer, & Graves, 2004) and is considered a general measure of cognitive flexibility and control (Uttl & Graf, 1997) or executive functioning (Moering et al., 2004). These abilities decline with age (Ivnik, Malec, Smith, Tangalos, & Petersen, 1996) and in dementia (Houx, Jolles, & Vreeling, 1993), which has made the Stroop test a popular test to evaluate various groups of patients with borderline or established brain pathology (Bohnen, Twijnstra, & Jolles, 1992; Davidson et al., 2003; Dulaney & Rogers, 1994; Graf, Uttl, & Tuokko, 1995; Houx et al., 1993; MacLeod, 1991; Uttl & Graf, 1997).

The popularity of the Stroop test in clinical and research settings (Lezak et al., 2004) means that it is important to determine the influence of age and age-extrinsic factors on test performance. Previous studies have provided inconclusive data about the effects of age, sex, and education on Stroop test performance. While most authors reported age-related decrements in Stroop test performance (Daigneault, Braun, & Whitaker, 1992; Feinstein, Brown, & Ron, 1994; Hamelers et al., 2000; Houx et al., 1993; Ivnik et al., 1996; Klein et al., 1997; Libon et al., 1994; Moering et al., 2004; Spreen & Strauss, 1998; Swerdlow, Filion, Geyer, & Braff, 1995; Van Boxtel, ten Tusscher, Metsemakers, Willems, & Jolles, 2001), Graf et al. (1995) did not find age to influence test performance. Sex differences in Stroop performance have been reported by some (Hameleers et al., 2000; Martin & Franzen, 1989; Moering et al., 2004; Van Boxtel et al., 2001) but not all (Houx et al., 1993; Klein et al., 1997; Swerdlow et al., 1995; Trenerry et al., 1989) authors. Again, education was found to be positively related to Stroop test performance by some authors (Houx et al., 1993; Moering et al., 2004; Van Boxtel et al., 2001; Hameleers et al., 2000), but not by others (Trenerry et al., 1989).

In the present study, we administered the Stroop test to a sample of 1,856 cognitively intact men and women aged 24 to 81 years with different levels of educational attainment, to establish normative data. In the past decade, there has been considerable debate about which methodological approach should be used

to derive normative data, i.e. a ‘traditional’ approach or a regression-based approach (see Fastenau, 1998; Fastenau & Adams, 1996; Heaton, Avitable, Grant, & Matthews, 1999; Heaton, Matthews, Grant, & Avitable, 1996; Van Breukelen & Vlaeyen, in press). The main difference between the two approaches is that with the traditional approach normative data are derived from raw test scores by splitting the sample into relevant demographic subgroups, whereas with the regression-based approach, normative data are derived from test scores as predicted from the relevant demographic variables. As both methods have their advantages, either methodological or in terms of ease of use, we established normative data using both methods.

Method

Participants

The data were derived from the Maastricht Aging Study (MAAS), a prospective study into the determinants of cognitive aging (Jolles, Houx, Van Boxtel, & Ponds, 1995). Participants in the MAAS study were recruited from the Registration Network Family Practices (*RegistratieNet Huisartspraktijken*, RNH), which includes 80,000 people who live in the province of Limburg in the Netherlands (Metsemakers, Höppener, Knottnerus, Kocken, & Limonard, 1992). RNH physicians classify health problems according to the International Classification of Health Problems in Primary Care (ICPC, Lamberts & Wood, 1987). The use of the RNH as a sample frame rather than a general population sample has the advantage that an eligible study sample could be selected beforehand. Thus, individuals with medical conditions known to interfere with cognition (i.e. cerebrovascular pathology, tumors of the nervous system, multiple sclerosis, epilepsy, parkinsonism, dementia, organic psychosis, schizophrenia, affective psychosis, and mental retardation) were identified using the RNH records and excluded from the sample frame. In total 10,396 individuals aged between 24 and 81 years were then randomly drawn from the RNH. They were informed about the study by their general practitioner rather than by the MAAS project staff, which was expected to have a facilitative effect on compliance, and were asked to return a prepaid postcard to indicate their willingness to participate in the study. In total, 4,490 people (43.2%) agreed to participate (3,531 individuals [34%] refused participation and 2,375 [22.8%] people did not return the postcard). Potential participants were then screened in a semi-structured interview to check for additional exclusion criteria that were not coded in the RNH (i.e. a history of transient ischemic attacks, brain surgery, hemodialysis for renal failure, electroconvulsive therapy, and chronic psychotropic drug use), which led to the exclusion of 301 individuals. Of the remaining 4,189 participants, 1,856 were randomly selected from 12 discontinuous age categories (25±1 years, 30±1 years, ..., 80±1 years) for participation in the study.

Not all data for the 1,856 participants administered the Stroop test were included in the analyses. The following exclusion criteria were used: a score below 24 on the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975), the occurrence of technical problems during test assessment, and more than twenty errors on the third Stroop subtask (indicative of possible cognitive problems). This led to the exclusion of data for 68 participants (data for 13, 52, and 3 participants were excluded for the above-mentioned reasons, respectively). Table 1 provides basic descriptive data for the sample. Level of Education (LE) was assessed by classifying formal schooling in a system often used in the Netherlands (De Bie, 1987), which is comparable to the International Standard Classification of Education (UNESCO, Paris, 1976). The participants were grouped as follows: those with at most primary education (LE low), those with junior vocational training (LE average), and those with senior vocational or academic training (LE high). The ethnic background of all participants was Caucasian, and all participants were native Dutch speakers.

Table 1. Descriptive characteristics of the sample (N=1,788).

Age		Age		Level of Education (frequency)			Male:Female
Group	N	Mean	SD	Low	Average	High	Ratio
25±1 year	159	25.40	.90	21	77	61	81:78
30±1 year	154	30.60	.87	22	79	53	75:79
35±1 year	157	35.55	.90	33	77	47	79:78
40±1 year	155	40.49	.87	37	72	46	78:77
45±1 year	162	45.54	.92	49	73	40	84:78
50±1 year	160	50.29	.88	52	71	37	81:79
55±1 year	157	55.46	.90	81	49	27	76:81
60±1 year	158	60.59	.84	73	65	19	76:82
65±1 year	153	65.45	.87	83	50	20	77:76
70±1 year	155	70.31	.82	80	57	18	74:81
75±1 year	158	75.16	.81	80	53	24	81:77
80±1 year	60	79.81	.82	38	9	12	29:31
Total	1,788	51.41	16.37	649	732	404	891:897

Note. Data on Level of Education were missing from 3 participants.

Procedure and Instruments

All participants were tested individually at the neuropsychological laboratory of the Brain & Behaviour Institute in Maastricht, the Netherlands, using the Stroop test version most commonly used in this country (Hammes, 1973). This Stroop test version consists of three subtasks. The stimulus material for each of these subtasks is shown on a white sheet of paper that is landscape oriented (A4 letter format, 11.69 inch x 8.26 inch [29.70 cm x 20.99 cm]). The 100 stimuli

for each subtask are distributed evenly in a 10 x 10 matrix on each sheet of paper with a margin of about 1.97 inch (5 cm) at the top, 0.59 inch (1.5 cm) on the left and on the right, and 1.57 inch (4 cm) at the bottom. The first subtask shows color words in random order (red, blue, yellow, green) printed in black ink (non-capital letters, 0.157 inch [0.4 cm] high). Subtask two displays solid color patches of 0.276 inch x 0.787 inch (0.7 cm x 2.0 cm) in one of these four basic colors. The third subtask contains color words printed in an incongruous ink color (non-capital letters, 0.157 inch [0.4 cm] high), e.g. the word yellow printed in red ink. The Dutch words for red, blue, yellow, and green are similar to their English equivalents in length and pronunciation time (i.e. *rood*, *blauw*, *geel*, and *groen*, respectively).

The participants were instructed to read the words, name the colors, and, finally, name the ink color of the printed words as quickly and as accurately as possible in the three subsequent subtasks. There was no time limit to complete a subtask. The times needed to complete each Stroop subtask served as dependent measures (Stroop I, Stroop II, and, Stroop III, respectively). An interference measure was calculated by subtracting the average time needed to complete the first two subtasks from the time needed to complete the third subtask (Interference = Stroop III - [(Stroop I + Stroop II) / 2]; Valentijn et al., 2005). The examiners did not point out errors made during the test. Although many participants spontaneously corrected themselves when they noticed an error (which requires a certain amount of time, so that the Stroop I, Stroop II and Stroop III scores were, to some extent, indirectly corrected for poor accuracy), this was not always the case. Therefore, the number of errors that were not self-corrected was also recorded for each Stroop subtask (Error I, Error II, and, Error III, respectively).

The data were collected by five examiners who had been intensively trained in test administration. Regular training sessions were scheduled to ensure uniform administration of the Stroop.

Data analysis

We first established which demographic variables were, and which were not, predictive for the different Stroop scores, so that the normative data could be appropriately corrected for these variables. Thus, the Stroop scores were regressed on Age, Age², Sex, Level of Education, and all two-way interactions. Age was centered (calendar age - 50) before computing the quadratic terms and interactions to avoid multicollinearity (Marquardt, 1980). Sex was dummy coded with male = 1 and female = 0. Level of Education (LE) was dummy coded with two dummies (LE low and LE high) and LE average as the reference category. The full models were then reduced in a stepwise way by eliminating the least significant predictor if its two-tailed *p*-value was higher than .005 (a relatively small *alpha* was taken to avoid Type I errors). Note that a predictor

was never removed from the model as long as it was also included in a higher-order term in the model. The reason for this is that the p -value of any predictor is arbitrary (depending on the coding used for the predictors) if that predictor is part of any higher-order predictor in the model (Aiken & West, 1991). The dummies LE low and LE high were always either both in, or both out of the model because they belong together and represent the effect of the categorical predictor LE. Similarly, their interactions with another predictor were always either both in, or both out of the model. The assumptions of regression analysis (homoscedasticity, normal distribution of the residuals, absence of multicollinearity and absence of 'influential cases') were tested for each model. The normal distribution of the residuals was investigated by inspection of the normal probability plots. The occurrence of multicollinearity was checked by calculating the Variance Inflation Factors (VIFs), which should not exceed 10 (Belsley, Kuh, & Welsch, 1980). Cook's distances were calculated to identify possible influential cases. Homoscedasticity was evaluated by grouping participants into quartiles of the predicted scores and applying the Levene test to the residuals.

Next, normative data were calculated using the traditional and the regression-based approach. For the traditional approach, Stroop normative statistics (mean, SD) were determined based on the observed data per relevant subgroup (as identified by the significant predictors in the regression models). These statistics can be used to convert the raw scores into Z-scores ($Z = - [\text{observed score} - \text{mean score}] / \text{SD}$)¹, but this conversion is appropriate only if the raw scores are normally distributed. If this was not the case, cumulative percentages for the observed values of the scores are provided.

Using the regression-based approach, the raw Stroop scores of a person are converted into standardized residuals in three steps. First, the predicted scores of a person are calculated; secondly, the residuals are calculated ($e_i = - [\text{observed score} - \text{predicted score}]$)²; and thirdly, the residuals are standardized ($Z_i = e_i / \text{SD}[\text{residual}]$). These standardized residuals can be interpreted via a Z-distribution table with cumulative probabilities, but this interpretation is appropriate only if the distribution of the standardized residuals of the sample is normal. More details regarding the regression-based normative method can be found elsewhere (Van Breukelen & Vlaeyen, in press). All analyses were performed using the SPSS 11.5 for Windows software package.

¹ A Z-score is usually calculated as $(\text{observed score} - \text{mean score}) / \text{SD}$ (so without reversing the positive/negative sign), because in general a higher/lower observed score compared to the mean score signifies a better/worse performance than expected, respectively. With regard to the Stroop test scores, however, a higher score means a worse performance. For this reason, the sign of the residual value is reversed.

² The sign of the residual value is reversed for the reason given in note 1.

Results

The final models are presented in Table 2. A model for the Error I score is not presented because none of the independent variables contributed to the prediction of this score. No significant influence of outliers was observed (maximum Cook's distance equaled .066). The VIFs of the predictors were at most 2.567, well below the cut-off value.

Table 3 shows the zero-order correlations between the raw Stroop scores and the demographic variables used in the regression models. These regression models (see Table 2) showed that the Stroop I, Stroop II, Stroop III, and Interference scores were significantly affected by linear and quadratic Age components. Females had better Stroop II, Stroop III, and Interference scores than males, but no sex differences were found for the Stroop I and the Error scores. Participants with a lower educational attainment scored worse than their higher educated counterparts on all the Stroop scores with the exception of the Error I score. The Error II score was influenced by level of education only. Significant Age x LE low interactions were found for the Interference and Error III scores. The predicted Interference and Error III scores are shown as a function of Age and Level of education for male participants in Figures 1 and 2 (for females, these plots were identical for the Error III score, and similar for the Interference score – with a constant value subtracted from the predicted scores, see Table 2). As shown, the difference in the predicted Interference and Error III scores between participants with a low educational attainment and participants with a higher educational attainment increased significantly as a function of age.

Traditional normative data

Table 4 provides the raw Stroop score descriptive statistics (mean, SD) and sample sizes per relevant subgroup (as identified by the regression models presented in Table 2; the continuous predictor Age was categorized to Age groups of 25 ± 1 years, 30 ± 1 years, ..., 80 ± 1 years). These statistics can be used to transform the raw scores of a person into Z-scores, but this transformation is only appropriate if the raw scores per relevant subgroup are normally distributed. This was the case for the Stroop I, Stroop II, Stroop III, and Interferences raw scores (all p 's of Kolmogorov-Smirnov $Z > .005$, see Table 4), but the Z-score transformation was not appropriate to evaluate a person's Error I, Error II, and Error III scores because of highly skewed distributions of the raw scores in the majority of the subgroups. Thus, for the Error scores, cumulative percentages representing the proportion of people who scored at or above a certain raw score were derived from the observed distribution in the sample (see Appendix, Table A1).

Note. Coding of the predictors: Age = calendar age - 50; Age² = (calendar age - 50)²; Sex: Male = 1, Female = 0; LE Low: Low Education = 1, Average or High Education = 0; LE High: High Education = 1, Low or Average Education = 0.
* $p < .005$; ** $p < .001$

Table 3. Zero-order correlations between the predictors used in the regression analyses and the Stroop measures.

	1	2	3	4	5	6	7	8	9	10	11
1: Age	1										
2: Sex	0.01	1									
3: LE low	0.31**	^a -0.10**	1								
4: LE high	-0.19**	^a 0.13**	^a -0.41**	1							
5: Stroop I	0.38**	0.02	0.34**	-0.22**	1						
6: Stroop II	0.43**	0.08**	0.32**	-0.21**	0.76**	1					
7: Stroop III	0.58**	0.06	0.41**	-0.23**	0.60**	0.74**	1				
8: Interference	0.55**	0.05	0.37**	-0.20**	0.38**	0.53**	0.96**	1			
9: Error I	0.04	-0.01	0.08**	-0.06	0.05	0.07*	0.08**	0.07*	1		
10: Error II	0.08**	0.01	0.10**	-0.09**	0.07*	0.14**	0.14**	0.12**	0.11**	1	
11: Error III	0.18**	0.03	0.19**	-0.09**	0.18**	0.19**	0.32**	0.32**	0.11**	0.17**	1

^a These correlations (between two dummy variables) correspond to ϕ -coefficients (Hays & Winkler, 1971)

* $p < .005$; ** $p < .001$

Figure 1. Predicted Stroop Interference scores for male participants, by Level of Education as a function of age.

Figure 2. Predicted Stroop Error III scores for males and females, by Level of Education as a function of age.

As an example of the traditional method, a 40-year-old man with an average level of education scored 40 on Stroop I, 60 on Stroop II, and 85 on Stroop III. The interference score was $35 (= 85 - [(40 + 60) / 2])$, which corresponds to a Z-score of $.22 (= [- (35 - 37.36) / 10.55])$ and a p -value of $.58$, which can be considered “normal”. This person scored 0 on Error I, 1 on Error II, and 1 on Error III. The cumulative percentage tables in the Appendix (Table A1) should be used to norm these scores rather than Table 4 for the reason described. Thus, the person’s Error I score was equal to or better than 100% of the Error I scores in the normative sample, the Error II score was equal to or better than 21% of the Error II scores in the normative sample, and the Error III score was equal to or better than 22.2% of the Error III scores in the normative sample. These scores can all be considered “normal”.

Regression-based normative data

The regression models, combined with the standard deviations of the residuals, provide normative data: after calculation of a person’s predicted scores by means of the regression models in Table 2, the residuals of each score are calculated ($e_i = - [\text{observed score} - \text{predicted score}]$) and standardized ($Z_i = e_i / \text{SD} [\text{residual}]$) using Table 5. The $\text{SD}(\text{residuals})$ were calculated per quartile of the predicted scores to account for heteroscedasticity, because the Levene test rejected the assumption that variance was homogeneous in all the models (all p

Table 4. Descriptive statistics for the Stroop measures stratified by their relevant predictors.

Measure	Sex	LE	Age group											
			25±1	30±1	35±1	40±1	45±1	50±1	55±1	60±1	65±1	70±1	75±1	80±1
Stroop I	Males & Females	Mean	43.23	43.61	44.47	45.94	43.30	44.82	45.43	48.61	47.64	47.16	51.52	51.46
		SD	6.03	6.78	6.71	9.97	7.19	6.66	7.14	8.91	7.80	7.14	8.91	7.54
		n	21	22	33	37	49	52	81	73	83	80	80	38
		Mean	39.49	41.22	38.59	40.78	41.29	42.68	42.76	42.13	44.51	44.95	47.91	45.78
		SD	4.76	6.20	5.28	5.55	6.94	6.98	6.46	6.85	7.55	7.07	7.40	5.16
		n	77	79	77	72	73	71	49	65	50	57	53	9
		Mean	38.61	39.56	38.96	38.16	39.48	40.47	38.59	43.50	42.92	44.04	44.00	47.98
		SD	6.44	6.25	5.57	4.77	7.09	6.90	5.47	5.88	7.64	7.67	4.67	8.02
		n	61	53	47	46	40	37	27	19	20	18	24	12
			25±1	30±1	35±1	40±1	45±1	50±1	55±1	60±1	65±1	70±1	75±1	80±1
Stroop II	Males	Mean	61.37	56.90	57.38	59.48	58.91	57.33	59.97	62.05	63.65	63.77	69.81	71.97
		SD	5.24	7.35	10.05	8.46	8.45	9.41	12.07	11.24	9.69	10.66	14.17	12.80
		n	8	9	22	21	23	17	33	31	32	35	32	20
		Mean	55.36	54.16	50.96	51.35	54.34	54.94	58.59	57.61	62.74	61.83	65.83	63.85
		SD	7.46	8.11	6.49	8.97	10.35	9.43	7.92	8.51	14.06	10.36	10.82	7.73
		n	37	41	25	24	32	43	27	38	29	34	27	4
		Mean	50.36	51.55	50.49	51.73	51.37	53.77	52.10	56.77	56.61	60.07	59.36	60.14
		SD	7.43	7.03	7.92	7.45	8.23	8.45	7.94	5.67	9.72	8.63	8.70	5.74
		n	33	29	31	32	23	19	21	12	15	12	18	6
		Mean	58.70	54.45	57.67	57.20	51.70	56.14	57.74	62.50	61.36	60.80	67.90	71.38
Females	Females	SD	8.47	7.74	10.06	13.84	7.58	9.19	8.57	8.82	10.11	8.13	11.30	11.82
		n	13	13	11	16	26	35	48	42	51	45	48	18
		Mean	49.68	51.17	48.84	52.13	51.53	52.39	56.34	53.89	57.08	56.72	63.80	65.71
		SD	8.27	6.31	6.84	8.30	7.84	7.23	7.10	7.52	8.44	10.01	8.43	11.06
		n	40	38	52	48	41	28	22	27	21	23	26	5
		Mean	49.66	49.41	50.86	44.91	49.88	55.69	48.62	56.55	58.41	54.58	56.34	61.96
		SD	8.72	9.47	4.52	5.62	6.51	14.29	11.04	7.71	10.75	11.80	9.02	7.63
		n	28	24	16	14	17	18	6	7	5	6	6	6
			25±1	30±1	35±1	40±1	45±1	50±1	55±1	60±1	65±1	70±1	75±1	80±1

(continued on the next page)

Stroop III												
	25±1	30±1	35±1	40±1	45±1	50±1	55±1	60±1	65±1	70±1	75±1	80±1
Mean	99.03	86.97	89.03	94.84	106.02	106.18	108.51	105.32	116.99	121.11	137.33	150.96
SD	9.48	12.02	15.63	18.36	22.69	22.94	25.61	19.61	30.14	32.68	26.73	27.85
<i>n</i>	8	9	22	21	23	17	33	31	32	35	32	20
Mean	84.44	84.22	78.57	82.80	86.74	88.07	95.94	93.78	100.81	106.64	117.64	129.68
SD	14.24	15.15	12.73	13.12	20.62	20.47	15.20	16.26	20.07	23.76	21.31	27.30
<i>n</i>	37	41	25	24	32	43	27	38	29	34	27	4
Mean	73.92	81.85	77.74	82.86	83.26	81.83	83.81	94.40	99.20	102.40	109.46	114.57
SD	11.47	13.66	15.23	17.23	16.94	12.13	17.38	19.48	18.90	15.11	24.25	16.91
<i>n</i>	33	29	31	32	23	19	21	12	15	12	18	6
Mean	92.55	89.16	90.32	91.13	82.68	91.90	100.17	111.74	109.83	112.19	131.97	150.17
SD	16.74	17.67	13.99	24.64	15.74	13.43	18.32	21.16	22.15	20.24	30.48	36.24
<i>n</i>	13	13	11	16	26	35	48	42	51	45	48	18
Mean	73.93	76.85	77.08	80.32	83.82	83.22	95.42	88.01	97.75	106.03	114.55	126.22
SD	11.00	12.66	14.15	11.75	13.89	17.12	14.58	13.77	25.37	26.45	19.39	21.86
<i>n</i>	40	38	52	48	41	28	22	27	21	23	26	5
Mean	75.75	74.52	80.68	69.63	74.76	86.23	79.72	88.11	91.76	95.68	108.54	125.61
SD	13.50	17.37	11.20	10.52	9.17	15.50	18.72	11.90	19.55	16.26	29.94	33.28
<i>n</i>	28	24	16	14	17	18	6	7	5	6	6	6
Interference												
	25±1	30±1	35±1	40±1	45±1	50±1	55±1	60±1	65±1	70±1	75±1	80±1
Mean	47.14	36.29	38.75	43.03	53.65	54.90	55.87	49.55	60.70	65.33	76.36	88.95
SD	8.34	8.53	12.86	16.59	18.59	19.37	18.49	17.17	25.11	28.49	22.91	29.22
<i>n</i>	8	9	22	21	23	17	33	31	32	35	32	20
Mean	36.16	36.17	33.06	37.36	38.85	39.35	45.02	43.12	46.88	52.73	60.92	74.44
SD	11.19	10.95	9.56	10.55	14.94	15.94	14.58	13.50	14.03	18.94	20.06	24.52
<i>n</i>	37	41	25	24	32	43	27	38	29	34	27	4
Mean	29.52	36.47	33.33	37.68	38.01	34.04	38.77	44.35	49.86	49.95	57.90	61.24
SD	7.52	10.97	11.56	15.30	13.00	11.58	13.77	17.72	13.55	12.73	19.60	17.55
<i>n</i>	33	29	31	32	23	19	21	12	15	12	18	6
Mean	41.33	40.42	37.97	38.55	36.30	41.52	49.09	56.51	55.73	58.59	72.47	89.06
SD	15.42	14.80	10.05	16.35	10.94	10.29	14.79	18.42	18.72	18.34	26.45	29.11
<i>n</i>	13	13	11	16	26	35	48	42	51	45	48	18
Mean	30.13	31.04	33.72	33.76	37.48	35.56	46.16	41.10	47.07	55.91	58.54	70.81
SD	7.43	10.42	10.64	10.95	10.73	14.34	12.22	11.19	23.10	20.89	16.41	14.78
<i>n</i>	40	38	52	48	41	28	22	27	21	23	26	5

Error I	Males & Females	High	Mean	31.34	29.81	35.15	28.64	29.84	38.85	35.03	37.94	40.88	47.16	58.03	69.91	
			SD	9.82	12.95	10.29	8.56	9.60	11.53	11.81	7.35	7.85	7.68	25.88	27.93	
			<i>n</i>	28	24	16	14	17	18	6	7	5	6	6	6	
			All Age groups													
Error II	All LE's	Mean	0.21 ^a													
		SD	0.593													
		<i>n</i>	1788													
		All Age groups														
	Low	Mean	0.46 ^a													
		SD	0.94													
		<i>n</i>	649													
		Mean	0.33 ^a													
	Average	SD	0.81													
		<i>n</i>	732													
		Mean	0.22 ^a													
		SD	0.61													
Error III	Males & Females	High	<i>n</i>	404												
			25±1	30±1	35±1	40±1	45±1	50±1	55±1	60±1	65±1	70±1	75±1	80±1		
			Mean	1.14	0.50	0.82 ^a	0.57 ^a	0.86 ^a	1.46 ^a	0.80 ^a	0.89 ^a	1.01 ^a	1.15 ^a	2.63 ^a	2.45	
			SD	1.65	0.67	1.18	1.28	1.61	2.04	1.44	1.45	1.87	2.03	4.46	3.45	
	Low	<i>n</i>	21	22	33	37	49	52	80	73	83	80	79	38		
		Mean	0.52 ^a	0.53 ^a	0.45 ^a	0.39 ^a	0.68 ^a	0.49 ^a	0.55 ^a	0.54 ^a	0.96 ^a	0.47 ^a	0.60 ^a	1.33		
		SD	0.84	1.02	0.91	0.85	1.26	0.89	1.50	0.99	2.37	0.89	1.10	1.58		
		<i>n</i>	77	79	77	72	73	71	49	65	49	57	53	9		
	Average	Mean	0.33 ^a	0.60 ^a	0.32 ^a	0.57 ^a	0.35 ^a	0.27 ^a	0.30 ^a	0.26 ^a	0.42 ^a	0.72	1.33	0.92		
		SD	0.89	1.81	0.69	1.26	0.80	0.56	0.78	0.73	0.84	1.32	2.55	2.02		
		<i>n</i>	61	53	47	46	40	37	27	19	19	18	24	12		

Note. LE = Level of education. The normative statistics for the scores that are stratified by LE were based on a total sample of 1,785 instead of 1,788 because data on LE were missing from 3 participants. Data on Error III scores were missing from 4 participants.

^a The mean and SD for this subgroup cannot be used to convert the raw scores into Z-scores because the raw scores in this subgroup were not normally distributed (*p*-value of the Kolmogorov-Smirnov *Z* < .005).

< .01). The normal probability plots showed normally distributed standardized residuals for the Stroop I, Stroop II, Stroop III, and Interference models when these SD(residuals) as a function of predicted scores were used. However, the use of SD(residuals) as a function of predicted scores did not solve the problem of skewed Error II and Error III standardized residual scores, which makes the regression-based normative approach inappropriate for these scores. For this reason, there are no SD(residual) values for these scores presented in Table 5.

As an example of the regression-based method, the Interference score of 35 of the 40-year-old man from the previous example is considered. The predicted interference score for this person was 35.676 ($= 36.066 + (.500 * -10) + (.016 * (-102)) + (3.010 * 1) + (8.505 * 0) + (-2.092 * 0) + (.167 * 0) + (-.019 * 0)$). So, the residual was .676 ($= - [35 - 35.676]$) and the standardized residual was .053 ($= .676 / 12.667$). This standardized residual corresponds to an exact p-value of .52 and can be considered “normal”. Because the steps that are required in this regression-based approach require some active calculation from the user of the normative data, we also calculated simplified normative tables to increase the user-friendliness of the regression-based norms (see Appendix, Tables B1 – B4). If an individual is not exactly 25, 30, ..., 80 years old, then the person’s age should be rounded up to the closest age given. If the performance of such an individual is borderline-normal (e.g. Z value ≈ -1.64), the exact Z values can be determined using Tables 2 and 5.

Discussion

In the present study, we first determined which factors affect Stroop performance, i.e. Age, Sex, Level of Education, and all possible two-way interactions between these predictors. All the speed-dependent Stroop test scores (Stroop I, Stroop II, Stroop III, and Interference) were profoundly affected by linear and quadratic age effects and LE (see Table 2). Although MacLeod (1991) suggested that sex only has a minor influence on Stroop test performance at any age, we found clear sex differences on the Stroop II, Stroop III, and the Interference scores, with women outperforming men over the entire age range studied (there were no Sex x Age interactions). With respect to the accuracy measures (Error I, Error II, and Error III), the influence of the demographic variables was less profound: the Error I score was not affected by any of the demographic variables, whereas although the Error II and Error III scores were affected to some extent by demographic variables, the proportion of explained variance of the models predicting these scores was low ($R^2 < .07$).

Table 5. Standard deviations of residuals for the Stroop I, Stroop II, Stroop III, and Interference scores.

Score	Predicted score	SD(residual)
Stroop I	≤ 40.209	5.961
	Between 40.210 and 43.353	6.400
	Between 43.354 and 46.059	7.217
	≥ 46.060	7.921
Stroop II	≤ 51.661	7.988
	Between 51.662 and 55.861	8.459
	Between 55.862 and 60.713	9.419
	≥ 60.714	10.587
Stroop III	≤ 79.988	13.963
	Between 79.989 and 92.862	16.367
	Between 92.863 and 108.585	19.506
	≥ 108.586	25.936
Interference	≤ 34.845	11.037
	Between 34.846 and 41.636	12.667
	Between 41.637 and 54.849	15.856
	≥ 54.850	22.472

The impact of a low LE as compared with an average LE was larger than the impact of a high LE on all Stroop measures – with the exception of the Error I score (see B values, Table 2). In addition to the main effect of LE on all the Stroop scores, there were significant Age x LE Low interactions on the Interference and Error III scores. In other words, the influence of educational attainment was independent of the influence of age on the Stroop I, II, III, and Error II scores, but the difference in performance between the low and average/high educated participants increased as a function of age on the Interference and Error III scores. As the interference score is generally regarded as a measure of executive functioning, this interaction suggests that executive functioning declines with age, and that the magnitude of decline is influenced by a person's educational attainment. This result is consistent with the cognitive reserve hypothesis, which claims that certain environmental factors make individuals less vulnerable to age-related cognitive decline and pathological brain processes (Stern et al., 2003). Examples of such factors are education (Dufouil, Alpérovitch, & Tzourio, 2003; Le Carret et al., 2003), mental stimulation at work (Bosma, Van Boxtel, Ponds, Houx, & Jolles, 2002), and leisure activity (Scarmeas, Levy, Tang, Manly, & Stern, 2001).

We derived normative Stroop data using a traditional and a regression-based approach because both methods have their advantages and disadvantages. The best approach for establishing normative data can only be determined by com-

paring the normative data obtained with the two methods with the “true” situation (when all scores of the entire population are known, for example by using a simulation study). The population distribution of scores is never known in normative studies, but based on the results of the present study, some considerations can be given regarding the issue of regression-based versus traditional norms.

With the traditional method, raw scores are often converted into Z-scores to evaluate a person’s performance. This method is appropriate if the mean and SD are accurately estimated (i.e. the mean and SD of the sample approximate the true population mean and SD), and if the raw test scores are normally distributed. A problem that is intrinsically related to the traditional approach is that calculation of these statistics usually requires the total sample to be divided into subgroups. Indeed, performance on many (neuro)psychological tests is influenced by independent variables such as age, sex, and education (Mitrushina et al., 1999). As a result, the total sample has to be subdivided to provide norms corrected for these variables, which dramatically decreases the sample size on which the normative statistics are calculated. For example, splitting a sample by age group (12 levels), sex (2 levels), and level of education (3 levels), reduces the sample size per subgroup to $1/72^{\text{th}}$ of the total sample size. Consequently, the normative statistics are less accurate when estimated for subgroups than for the total sample (Van Breukelen & Vlaeyen, in press). The problem of dividing a sample into subgroups is especially pronounced if the subgroups are not balanced in size. This is usually the case in normative studies involving participants from a broad age range that are randomly selected. Indeed, if there is adequate random sampling, certain population trends should be reflected in the sample, i.e. the decrease in level of education as a function of increasing age (many elderly people educated before 1960 have a low level of educational attainment because they left school early or did not go to university or other institutes of higher education; Jolles et al., 1995). As a result, the sample sizes for certain subgroups are unusually small (e.g. for high educated elderly and low educated younger people), which yields very unreliable normative data for these subgroups – with the mean scores and SDs moving up, down, and up again across age groups due to chance trends (see Table 4). A possible solution for this problem is the use of broader age categories to increase the number of participants per subgroup, but then there is a problem with the boundary value. For example, if the age groups are as broad as 24-36 years, 39-51 years, and so on, the scores of people aged 36 and 39 (differing only 3 years) would be evaluated against different normative data whereas the scores of people aged 24 to 36 (differing 12 years) would be evaluated against the same normative data, which is not acceptable (Capitani, 1997).

In the regression-based approach, the problem of unreliable normative data for certain subgroups because the sample was subdivided does not occur. Indeed, regression-based norms provide more accurate estimates of population statistics because they are based on equations that are derived using the data for all

demographic groups (Van Breukelen & Vlaeyen, in press; Zachary & Gorsuch, 1985). In fact, normative data can even be provided for people with certain demographic characteristics that were not in the sample. For example, in the present study, discontinuous age groups were used (24-26 years, 29-31 years, 34-36 years, etc), but normative data for non-tested people aged e.g. 28 years could be determined by using the regression model and the SD(residual) derived from the scores of the tested people. Also, the problem of unbalanced data does not occur in the regression-based approach because any imbalance in the sample does not bias the estimation or testing of the regression weights but only causes some loss of statistical power (because the standard error of a regression weight is proportional to the $\sqrt{\text{VIF}}$ of the predictor at hand; see e.g. Kleinbaum, Kupper, Muller, & Nizam, 1998). In addition, the regression-based normative approach offers some interesting possibilities that cannot be achieved by using a traditional approach. For example, as an alternative for the Interference score, a regression model can be constructed that predicts the Stroop III score based on the Stroop I and Stroop II scores in addition to the relevant demographic variables. Such an approach avoids the use of difference scores such as the Interference score, which may reduce the problems that are associated with such scores, i.e. their typically low reliability. We did not provide such measures here, because normative tables for each relevant combination of predictors (Stroop I score, Stroop II score, age, sex, and/or level of education) would be too numerous. Thus, for such norms to be used, the three-step procedure to convert raw scores into standardized residuals would be required, which is not very user-friendly. A solution is to use a computer-based algorithm that performs these calculations automatically (which can easily be done with standard programs such as Microsoft Excel), or to use categorized values instead of continuous values.

Although the regression-based approach has some significant advantages over the traditional method, it requires some distributional assumptions that are not always satisfied. Indeed, in the present study, the regression-based method was appropriate to derive norms for the speed-dependent scores (Stroop I, Stroop II, Stroop III, and Interference scores), but there were problems with the accuracy measures (Error scores). An important difference between both types of scores is that the speed-dependent scores are continuous variables, whilst the accuracy measures are discrete. In general, linear regression analysis is primarily suitable to analyze continuous dependent variables, although in most cases discrete dependent variables can also be analyzed with regression analysis. For example, the regression-based normative approach was found to be appropriate in previous normative research with the Verbal Learning Test of Rey (Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2005), a test for declarative memory of which most of the scores are discrete variables (with a range of the scores of 0 to 15). The Stroop Error scores are also discrete variables that cover about the same range of scores, but a problem with the Error scores is that they are confined to a small part of the possible range of the scores and that they are highly skewed. When such data are analyzed with regression analysis, the assumption

of normally distributed residuals is usually violated (Fox, 1997). This assumption is arbitrary for most purposes – because regression analysis is robust against violations of the normality of the residuals assumption (Fox, 1997) – but in the context of the regression-based normative approach this assumption is important because the standardized residuals can only be interpreted via a Z-distribution table if they are normally distributed. However, heavily skewed scores that are confined to a small part of the possible range of scores can be normed by using a method similar to the regression-based approach, i.e. by dichotomizing the scores and analyzing them by logistic regression (Fox, 1997), but this method is more complex and beyond the scope of the present paper. Rather than using logistic regression, we used cumulative percentages (see Appendix Table A1) to evaluate the Error scores of a testee. The use of cumulative percentages avoids the problem of non-normally distributed raw scores (which is required to use the Z-score transformation in the traditional approach), but the problems associated with the division of the sample into subgroups remain. Indeed, for the Error III score, the sample had to be divided into many subgroups because of significant effects of Age and LE on this score, which means that the cumulative percentages of the Error III scores for certain subgroups are derived from the data of small samples (see the *ns* in Table 4). Thus, caution is needed when using the normative cumulative percentages for the Error III score.

Although the normative data presented here are for Dutch-speaking participants, the Stroop test has been found to be culturally robust in a large-scale study that involved seven different languages (English, Danish, Dutch, French, German, Greek, and Spanish; Møller et al., 1998). This suggests that the normative data can be used for people living in different cultural areas or for people who speak other languages, but more research is needed. Houx et al. (2002) showed that the shortened Stroop Test (with 40 items per subtest instead of 100) was a reliable test (test/re-test correlation with 2-week test-interval of .80 for the 40-item Stroop III score) in a large sample of elderly participants in Great Britain, Ireland, and the Netherlands ($N=5,804$; age range 70-82 years). The Pearson's correlations varied less than .05 for each of the three countries, which suggest that the psychometric properties of the Stroop Test are similar across countries and languages. As a final remark, it should be kept in mind we used the Hammes (1973) version of the Stroop test. Although the Stroop interference effect is a robust effect that has been observed with numerous different versions of the Stroop test, more research is needed regarding the comparability and interchangeability of the different Stroop test versions.

Appendix

Table A1. Cumulative percentages for the raw Error I, Error II and Error III scores that represent the proportion of individuals who score at or above 0, 1, 2, 3, 4, and 5, stratified by their relevant predictors.

Error I	All LE's		MALE & FEMALE											
		Raw score	All Age Groups											
		5	.2											
		4	.6											
		3	1.4											
		2	3.8											
		1	14.7											
		0	100.0											
Error II	LE low		MALE & FEMALE											
		Raw score	All Age Groups											
		5	.6											
		4	1.5											
		3	4.0											
		2	11.6											
		1	27.2											
		0	100.0											
	LE average	Raw score	All Age Groups											
		5	.5											
		4	.8											
		3	2.2											
		2	7.1											
		1	21.0											
		0	100.0											
	LE high	Raw score	All Age Groups											
		5	.2											
		4	1.5											
		3	1.5											
		2	5.4											
		1	14.1											
		0	100.0											
Error III	LE low		MALE & FEMALE											
		Raw score	Age group											
			25±1	30±1	35±1	40±1	45±1	50±1	55±1	60±1	65±1	70±1	75±1	80±1
		5	4.8	0.0	0.0	2.7	2.0	9.6	1.3	4.1	6.0	7.5	17.7	18.4
		4	9.5	0.0	3.0	5.4	4.1	17.3	3.8	4.1	9.6	11.3	22.8	21.1
		3	19.0	0.0	12.1	8.1	12.2	25.0	11.3	8.2	14.5	15.0	27.8	28.9
		2	28.6	9.1	27.3	10.8	22.4	34.6	22.5	24.7	24.1	27.5	38.0	39.5
		1	47.6	40.9	39.4	27.0	36.7	48.1	36.3	42.5	37.3	42.5	55.7	63.2
		0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	LE average	Raw score	Age group											
			25±1	30±1	35±1	40±1	45±1	50±1	55±1	60±1	65±1	70±1	75±1	80±1
		5	0.0	0.0	0.0	0.0	1.4	0.0	4.1	1.5	2.0	0.0	1.9	0.0
		4	0.0	2.5	2.6	1.4	4.1	1.4	4.1	1.5	2.0	1.8	3.8	22.0
		3	5.2	8.9	5.2	4.2	13.7	5.6	8.2	3.1	8.2	5.3	5.7	22.0
		2	11.7	13.9	10.4	11.1	21.9	11.3	16.3	10.8	18.4	10.5	17.0	22.0
		1	35.1	27.8	27.3	22.2	27.4	31.0	16.3	35.4	42.9	29.8	32.1	66.7
		0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	LE high	Raw score	Age group											
			25±1	30±1	35±1	40±1	45±1	50±1	55±1	60±1	65±1	70±1	75±1	80±1
		5	1.6	3.8	0.0	4.3	0.0	0.0	0.0	0.0	0.0	5.6	12.5	8.3
		4	1.6	3.8	0.0	6.5	0.0	0.0	0.0	0.0	0.0	5.6	12.5	8.3
		3	3.3	5.7	2.1	8.7	5.0	0.0	3.7	5.3	0.0	5.6	12.5	8.3
		2	9.8	5.7	8.5	10.9	10.0	5.4	11.1	5.3	21.1	22.2	25.0	16.7
		1	16.4	28.3	21.3	26.1	20.0	21.6	14.8	15.8	21.1	33.3	37.5	33.3
		0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Note. LE = Level of education.

Table B1. Regression based normative data for the Stroop I score stratified by Age (25, 30, ..., 80 years) and Level of Education. The raw score leading to a particular Z-value is given for Z-values indicating the percentiles 5, 10, 20, 50, 80, 90 and 95.

	Z value	Cum. prob.	MALE & FEMALE											
			Age in years											
			25	30	35	40	45	50	55	60	65	70	75	80
LE low	1.64	.95	31.9	31.9	32.0	32.3	32.7	33.3	34.0	33.7	34.8	35.9	37.3	38.8
	1.28	.90	34.5	34.5	34.6	34.9	35.3	35.9	36.6	36.6	37.6	38.8	40.1	41.6
	0.84	.80	37.6	37.6	37.8	38.0	38.5	39.0	39.8	40.1	41.1	42.3	43.6	45.1
	0	.50	43.7	43.7	43.8	44.1	44.5	45.1	45.8	46.7	47.8	48.9	50.3	51.7
	-0.84	.20	49.8	49.8	49.9	50.2	50.6	51.2	51.9	53.4	54.4	55.6	56.9	58.4
	-1.28	.10	52.9	52.9	53.1	53.3	53.8	54.3	55.1	56.9	57.9	59.1	60.4	61.9
	-1.64	.05	55.5	55.5	55.7	55.9	56.4	56.9	57.7	59.7	60.7	61.9	63.3	64.7
LE average	Z value	Cum. prob.	Age in years											
			25	30	35	40	45	50	55	60	65	70	75	80
	1.64	.95	30.3	30.3	29.7	30.0	30.4	31.0	31.8	32.6	32.3	33.5	33.7	35.2
	1.28	.90	32.5	32.5	32.0	32.3	32.7	33.3	34.1	34.9	34.9	36.1	36.5	38.0
	0.84	.80	35.1	35.1	34.9	35.1	35.6	36.1	36.9	37.8	38.1	39.3	40.0	41.5
	0	.50	40.1	40.1	40.2	40.5	40.9	41.5	42.2	43.1	44.2	45.3	46.7	48.1
	-0.84	.20	45.1	45.1	45.6	45.9	46.3	46.9	47.6	48.5	50.2	51.4	53.3	54.8
LE high	-1.28	.10	47.7	47.7	48.4	48.7	49.1	49.7	50.4	51.3	53.4	54.6	56.8	58.3
	-1.64	.05	49.9	49.9	50.7	51.0	51.4	52.0	52.7	53.6	56.0	57.2	59.7	61.1
	Z value	Cum. prob.	Age in years											
			25	30	35	40	45	50	55	60	65	70	75	80
	1.64	.95	28.8	28.8	28.9	29.2	29.7	30.2	30.2	31.1	32.2	32.0	33.3	33.6
	1.28	.90	31.0	31.0	31.1	31.4	31.8	32.4	32.5	33.4	34.5	34.6	35.9	36.5
	0.84	.80	33.6	33.6	33.7	34.0	34.4	35.0	35.4	36.2	37.3	37.8	39.1	40.0
	0	.50	38.6	38.6	38.7	39.0	39.4	40.0	40.7	41.6	42.7	43.8	45.2	46.6
	-0.84	.20	43.6	43.6	43.7	44.0	44.4	45.0	46.1	47.0	48.0	49.9	51.2	53.3
	-1.28	.10	46.2	46.2	46.4	46.6	47.1	47.6	48.9	49.8	50.8	53.1	54.4	56.8
	-1.64	.05	48.4	48.4	48.5	48.8	49.2	49.8	51.2	52.1	53.1	55.7	57.0	59.6

Note. LE = Level of education; Cum. prob. = cumulative probability.

Table B2. Regression based normative data for the Stroop II score stratified by Age (25, 30, ..., 80 years), Sex, and Level of Education. The raw score leading to a particular Z-value is given for Z-values indicating the percentiles 5, 10, 20, 50, 80, 90 and 95.

Z	Cum. prob.	MALE										FEMALE														
		Age in years										Age in years														
		25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80	
LE low	1.64	.95	42.8	42.3	42.1	42.3	42.8	43.6	44.9	44.5	46.4	48.7	51.3	54.3	42.0	41.5	41.3	41.4	42.0	41.3	42.5	44.0	44.1	46.3	48.9	51.9
	1.28	.90	46.2	45.7	45.5	45.6	46.2	47.0	48.3	48.3	50.3	52.5	55.1	58.1	45.0	44.5	44.3	44.5	45.0	44.6	45.9	47.4	47.9	50.1	52.8	55.7
	0.84	80	50.3	49.8	49.6	49.8	50.3	51.2	52.4	53.0	54.9	57.2	59.8	62.8	48.7	48.2	48.0	48.2	48.7	48.8	50.0	51.6	52.5	54.8	57.4	60.4
	0	.50	58.2	57.7	57.5	57.7	58.2	59.1	60.3	61.9	63.8	66.1	68.7	71.7	55.9	55.3	55.1	55.3	55.8	56.7	57.9	59.5	61.4	63.7	66.3	69.3
	-0.84	.20	66.2	65.6	65.4	65.6	66.1	67.0	68.2	70.8	72.7	75.0	77.6	80.6	63.0	62.4	62.2	62.4	62.9	64.6	65.8	67.4	70.3	72.6	75.2	78.2
	-1.28	.10	70.3	69.8	69.6	69.8	70.3	71.1	72.4	75.4	77.4	79.6	82.2	85.2	66.7	66.2	66.0	66.1	66.7	68.8	70.0	71.5	75.0	77.2	79.9	82.8
-1.64	.05	73.7	73.2	73.0	73.2	73.7	74.5	75.8	79.2	81.2	83.4	86.1	89.0	69.7	69.2	69.0	69.2	69.7	72.2	73.4	74.9	78.8	81.0	83.7	86.6	
LE average	Z	Cum. prob.	25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
	1.64	.95	40.1	39.6	39.4	39.6	40.1	41.0	40.6	42.2	44.1	44.5	47.1	50.1	38.5	38.0	37.8	38.0	38.5	38.6	39.8	41.4	41.7	44.0	44.7	47.7
	1.28	.90	43.2	42.7	42.5	42.6	43.2	44.0	44.0	45.6	47.5	48.3	50.9	53.9	41.4	40.9	40.7	40.9	41.4	41.6	42.9	44.4	45.1	47.4	48.5	51.5
	0.84	.80	46.9	46.4	46.2	46.4	46.9	47.8	48.2	49.7	51.7	52.9	55.6	58.5	44.9	44.4	44.2	44.4	44.9	45.4	46.6	48.2	49.3	51.5	53.2	56.1
	0	.50	54.0	53.5	53.3	53.5	54.0	54.9	56.1	57.6	59.6	61.8	64.5	67.4	51.6	51.1	50.9	51.1	51.6	52.5	53.7	55.3	57.2	59.4	62.1	65.0
	-0.84	.20	61.1	60.6	60.4	60.6	61.1	62.0	64.0	65.6	67.5	70.7	73.4	76.3	58.3	57.8	57.6	57.8	58.3	59.6	60.8	62.4	65.1	67.4	71.0	73.9
-1.28	.10	64.8	64.3	64.1	64.3	64.8	65.7	68.1	69.7	71.6	75.4	78.0	81.0	61.8	61.3	61.1	61.3	61.8	63.3	64.5	66.1	69.2	71.5	75.6	78.6	
-1.64	.05	67.9	67.4	67.2	67.3	67.9	68.7	71.5	73.1	75.0	79.2	81.8	84.8	64.7	64.2	64.0	64.2	64.7	66.3	67.6	69.1	72.6	74.9	79.4	82.4	
LE high	Z	Cum. prob.	25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
	1.64	.95	37.8	38.0	37.9	38.0	38.5	38.6	39.9	41.4	41.8	44.0	44.7	47.7	36.2	35.6	35.5	35.6	36.2	37.0	38.2	39.0	41.0	41.7	44.3	45.3
	1.28	.90	40.8	40.9	40.7	40.9	41.4	41.7	42.9	44.5	45.2	47.4	48.6	51.5	39.0	38.5	38.3	38.5	39.0	39.9	41.1	42.1	44.0	45.0	47.7	49.1
	0.84	.80	44.6	44.4	44.2	44.4	44.9	45.4	46.6	48.2	49.3	51.6	53.2	56.2	42.6	42.0	41.9	42.0	42.5	43.4	44.6	45.8	47.7	49.2	51.8	53.8
	0	.50	51.7	51.1	51.0	51.1	51.6	52.5	53.7	55.3	57.2	59.5	62.1	65.1	49.3	48.7	48.6	48.7	49.3	50.1	51.3	52.9	54.8	57.1	59.7	62.7
	-0.84	.20	58.8	57.8	57.7	57.8	58.4	59.6	60.8	62.4	65.1	67.4	71.0	74.0	56.0	55.5	55.3	55.4	56.0	56.8	58.1	60.0	61.9	65.0	67.6	71.6
-1.28	.10	62.5	61.4	61.2	61.3	61.9	63.3	64.6	66.1	69.3	71.5	75.7	78.6	59.5	59.0	58.8	59.0	59.5	60.3	61.6	63.7	65.7	69.2	71.8	76.2	
-1.64	.05	65.5	64.2	64.1	64.2	64.7	66.4	67.6	69.2	72.7	74.9	79.5	82.4	62.4	61.8	61.7	61.8	62.4	63.2	64.4	66.8	68.7	72.5	75.2	80.1	

Note. LE = Level of education; Cum. prob. = cumulative probability.

Table B3. Regression based normative data for the Stroop III score stratified by Age (25, 30, ..., 80 years), Sex, and Level of Education. The raw score leading to a particular Z-value is given for Z-values indicating the percentiles 5, 10, 20, 50, 80, 90 and 95.

LE low	Z	Cum. prob.	MALE															FEMALE														
			Age in years															Age in years														
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80						
1.64	.95	64.9	63.3	62.8	63.5	65.4	68.4	72.5	67.3	73.7	81.3	90.0	99.9	65.6	64.0	63.5	64.2	60.9	63.9	68.0	73.3	69.2	76.8	85.6	95.5							
1.28	.90	71.9	70.3	69.9	70.5	72.4	75.4	79.5	76.6	83.0	90.6	99.4	109.3	71.5	69.9	69.4	70.1	67.9	70.9	75.1	80.4	78.6	86.2	94.9	104.8							
0.84	.80	80.5	78.9	78.4	79.1	81.0	84.0	88.1	88.0	94.5	102.0	110.8	120.7	78.7	77.1	76.6	77.3	76.5	79.5	83.6	88.9	90.0	97.6	106.3	116.2							
0	.50	86.9	85.3	84.8	85.5	87.4	90.4	94.5	94.4	100.9	108.4	117.2	127.1	84.2	82.6	82.1	82.8	81.9	84.9	89.0	93.1	96.2	103.7	112.4	122.3							
-0.84	.20	113.3	111.7	111.2	111.9	113.7	116.7	120.9	131.6	138.0	145.6	154.4	164.3	106.2	104.6	104.1	104.8	109.3	112.3	116.4	121.7	133.6	141.2	149.9	159.8							
-1.28	.10	121.8	120.2	119.8	120.5	122.3	125.3	129.5	143.0	149.4	157.0	165.8	175.7	113.4	111.8	111.3	112.0	117.9	120.9	125.0	130.3	145.0	152.6	161.3	171.2							
-1.64	.05	128.9	127.3	126.8	127.5	129.4	132.3	136.5	152.3	158.8	166.4	175.1	185.0	119.3	117.6	117.2	117.9	124.9	127.9	132.0	137.3	154.3	161.9	170.6	180.5							
LE average	Z	Cum. prob.	Age in years															Age in years														
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80						
			1.64	.95	56.8	55.1	54.7	55.4	57.2	60.2	64.4	64.5	71.0	68.0	76.8	86.7	56.2	54.6	54.2	54.9	56.7	55.8	59.9	65.2	66.5	74.1	72.3	82.2				
			1.28	.90	62.6	61.0	60.6	61.3	63.1	66.1	70.3	71.5	78.0	77.4	86.1	96.0	61.3	59.6	59.2	59.9	61.7	61.7	65.8	71.1	73.5	81.1	81.6	91.5				
			0.84	.80	69.8	68.2	67.8	68.5	70.3	73.3	77.5	80.1	86.6	88.8	97.5	107.4	67.4	65.8	65.3	66.0	67.9	68.9	73.0	78.3	82.1	89.7	93.0	102.9				
			0	.50	83.6	82.0	81.5	82.2	84.1	87.1	91.2	96.5	103.3	110.6	119.3	129.2	70.7	69.2	68.7	69.4	71.3	74.6	78.6	82.6	86.7	92.0	98.5	106.1	114.8	124.7		
			-0.84	.20	97.3	95.7	95.3	96.0	97.8	100.8	105.0	112.9	119.3	132.3	141.1	151.0	90.9	89.2	88.8	89.5	91.3	96.3	100.5	105.8	114.9	122.5	136.6	146.5				
-1.28	.10	104.5	102.9	102.5	103.2	105.0	108.0	112.2	121.5	127.9	143.7	152.5	162.4	97.0	95.4	94.9	95.6	97.5	103.6	107.7	113.0	123.5	131.0	148.0	157.9							
-1.64	.05	110.4	108.8	108.4	109.1	110.9	113.9	118.1	128.5	134.9	153.1	161.8	171.7	102.0	100.4	100.0	100.7	102.5	109.4	113.6	118.9	130.5	138.1	157.4	167.3							
LE high	Z	Cum. prob.	Age in years															Age in years														
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80						
			1.64	.95	56.8	55.2	54.8	55.5	57.4	60.4	64.6	64.7	71.2	68.2	77.0	86.9	56.2	54.6	54.3	55.0	56.8	55.8	56.0	61.3	62.6	70.2	68.4	78.3				
			1.28	.90	61.9	60.2	59.8	60.5	59.3	62.2	66.4	71.7	74.1	81.7	82.2	92.1	57.4	55.8	55.3	56.0	57.9	60.9	61.9	67.2	69.6	77.2	77.8	87.6				
			0.84	.80	68.0	66.4	65.9	66.6	66.5	69.4	73.6	78.9	82.7	90.3	93.6	103.5	63.5	61.9	61.5	62.2	64.0	67.0	69.1	74.4	78.2	85.8	89.2	99.1				
			0	.50	79.7	78.1	77.7	78.4	80.2	83.4	87.3	92.6	99.1	106.7	115.4	125.3	75.3	73.6	73.2	73.9	75.7	78.7	82.9	88.2	94.6	102.2	111.0	120.8				
			-0.84	.20	91.5	89.8	89.4	90.1	94.0	96.9	101.1	106.4	115.5	123.1	137.2	147.1	87.0	85.4	84.9	85.6	87.5	90.5	96.6	101.9	111.0	118.6	132.7	142.6				
-1.28	.10	97.6	96.0	95.5	96.2	101.2	104.1	108.3	113.6	124.1	131.6	148.6	158.5	93.1	91.5	91.1	91.8	93.6	96.6	103.7	109.1	119.6	127.2	144.2	154.0							
-1.64	.05	102.6	101.0	100.6	101.3	107.0	110.0	114.2	119.5	131.1	138.7	158.0	167.9	98.2	96.5	96.1	96.8	98.6	101.6	109.7	115.0	126.6	137.2	153.5	163.4							

Table B4. Regression based normative data for the Stroop Interference score stratified by Age (25, 30, ..., 80 years), Sex, and Level of Education. The raw score leading to a particular Z-value is given for Z-values indicating the percentiles 5, 10, 20, 50, 80, 90 and 95.

LE low	Z value	Cum. prob.	MALE										FEMALE													
			Age in years										Age in years													
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
	1.64	.95	20.1	19.9	20.4	16.5	18.6	21.6	25.3	19.0	24.3	30.5	37.4	45.1	17.1	16.9	17.4	18.7	20.9	18.6	22.3	26.8	21.3	27.5	34.4	42.1
	1.28	.90	24.7	24.4	25.0	22.2	24.4	27.3	31.0	27.1	32.4	38.6	45.5	53.2	21.7	21.4	22.0	23.3	25.4	24.3	28.0	32.5	29.4	35.5	42.5	50.2
	0.84	.80	30.3	30.0	30.5	29.2	31.3	34.3	38.0	37.0	42.3	48.4	55.4	63.1	27.3	27.0	27.5	28.9	31.0	31.3	35.0	39.5	39.3	45.4	52.4	60.1
	0	.50	40.9	40.6	41.2	42.5	44.6	47.6	51.3	55.9	61.2	67.3	74.3	82.0	37.9	37.6	38.2	39.5	41.6	44.6	48.3	52.8	58.2	64.3	71.2	79.0
	-0.84	.20	51.5	51.3	51.8	55.8	58.0	60.9	64.6	74.7	80.1	86.2	93.1	100.9	48.5	48.3	48.8	50.1	52.3	57.9	61.6	66.2	77.1	83.2	90.1	97.9
	-1.28	.10	57.1	56.9	57.4	62.8	64.9	67.9	71.6	84.6	90.0	96.1	103.0	110.8	54.1	53.8	54.4	55.7	57.8	64.9	68.6	73.1	86.9	93.1	100.0	107.7
	-1.64	.05	61.7	61.4	61.9	68.5	70.6	73.6	77.3	92.7	98.0	104.2	111.1	118.8	58.7	58.4	58.9	60.3	62.4	70.6	74.3	78.8	95.0	101.2	108.1	115.8
LE average	Z value	Cum. prob.	Age in years										Age in years													
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
			15.8	14.7	14.4	14.9	16.2	18.3	16.0	19.7	24.2	18.6	24.7	31.6	15.5	14.4	14.1	14.6	15.9	15.3	18.2	16.7	21.2	26.5	21.7	28.6
	1.64	.95	20.4	19.3	19.0	19.5	20.8	22.9	21.7	25.4	29.9	26.7	32.8	39.7	19.4	18.3	18.0	18.5	19.8	19.9	22.8	22.4	26.9	32.2	29.8	36.7
	0.84	.80	25.9	24.8	24.5	25.0	26.3	28.4	28.7	32.4	36.9	36.6	42.7	49.6	24.3	23.2	22.9	23.4	24.7	25.4	28.3	29.3	33.8	39.1	39.7	46.6
	0	.50	36.6	35.5	35.2	35.7	37.0	39.1	42.0	45.7	50.2	55.5	61.6	68.5	33.6	32.5	32.2	32.7	34.0	36.1	39.0	42.7	47.2	52.5	58.6	65.5
	-0.84	.20	47.2	46.1	45.8	46.3	47.6	49.7	55.3	59.0	63.5	74.4	80.5	87.4	42.8	41.7	41.4	41.9	43.2	46.7	49.6	56.0	60.5	65.8	77.4	84.3
	-1.28	.10	52.8	51.7	51.4	51.9	53.2	55.3	62.3	66.0	70.5	84.2	90.3	97.2	47.7	46.6	46.3	46.8	48.1	52.3	55.2	63.0	67.5	72.8	87.3	94.2
	-1.64	.05	57.3	56.2	55.9	56.4	57.7	59.8	68.0	71.7	76.2	92.3	98.4	105.3	51.7	50.6	50.3	50.8	52.1	56.8	59.7	68.7	73.2	78.5	95.4	102.3
LE high	Z value	Cum. prob.	Age in years										Age in years													
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
			14.2	15.7	15.3	15.7	14.2	16.2	19.0	17.4	21.8	27.0	22.2	29.0	13.8	12.7	12.3	12.7	13.9	15.9	16.0	19.6	18.8	24.0	19.1	25.9
	1.64	.95	18.7	19.6	19.2	19.6	18.8	20.8	23.6	23.1	27.5	32.7	30.2	37.0	17.8	16.6	16.2	16.6	17.8	19.8	20.6	24.2	24.5	29.7	27.2	34.0
	0.84	.80	24.3	24.5	24.1	24.5	24.3	26.3	29.1	30.1	34.5	39.7	40.1	46.9	22.7	21.5	21.1	21.5	22.7	24.7	26.1	29.7	31.5	36.7	37.1	43.9
	0	.50	35.0	33.8	33.4	33.8	35.0	37.0	39.8	43.4	47.8	53.0	59.0	65.8	31.9	30.8	30.4	30.8	32.0	34.0	36.8	40.4	44.8	50.0	56.0	62.8
	-0.84	.20	45.6	43.0	42.6	43.0	45.6	47.6	50.4	56.7	61.1	66.3	77.9	84.7	41.2	40.0	39.6	40.0	41.2	43.2	47.4	51.0	58.1	63.3	74.9	81.7
	-1.28	.10	51.2	47.9	47.5	47.9	51.2	53.2	56.0	63.7	68.1	73.3	87.8	94.6	46.1	44.9	44.5	44.9	46.1	48.1	53.0	56.6	65.1	70.3	84.8	91.6
	-1.64	.05	55.7	51.9	51.5	51.9	55.8	57.8	60.6	69.4	73.8	79.0	95.9	102.7	50.0	48.9	48.5	48.9	50.1	52.1	57.6	61.2	70.8	76.0	92.9	99.7

Note. LE = Level of education; Cum. prob. = cumulative probability.

Chapter 3

Rey's Verbal Learning Test: normative data for 1,855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation

Wim Van der Elst, Martin P. J. Van Boxtel, Gerard J. P. Van Breukelen & Jelle Jolles

Abstract

The Verbal Learning Test (VLT; Rey, 1958) evaluates the declarative memory. Despite its extensive use, it has been difficult to establish normative data because test administration has not been uniform. The purpose of the present study was to gather normative data for the VLT for a large number ($N=1,855$) of healthy participants aged 24 to 81 years, using a procedure in which the words to be learned were presented either verbally or visually.

The results showed that VLT performance decreased in an age-dependent manner from an early age. The learning capacity of younger versus older adults differed quantitatively rather than qualitatively. Females and higher educated participants outperformed males and lower educated participants over the entire age range tested. Presentation mode affected VLT performance differently: auditory presentation resulted in a better recall on Trial 1 (a short-term or working memory measure), whereas visual presentation yielded a better performance on Trial 3, Trial 4, and Delta (a learning measure).

Introduction

The declarative aspect of memory – the memory for specific facts or experiences (Squire, 1987) – is of particular interest in neuropsychology (Lezak, 1995; Woodruff-Pak, 1997). Consequently, a large number of tests to assess declarative memory have been developed, most notably the Verbal Learning Test of Rey (1958), the Buschke Selective Reminding Test (Buschke, 1973), and the California Verbal Learning Test (Delis et al., 1987). The present study focuses on the Verbal Learning Test (VLT; Brand & Jolles, 1985; Rey, 1958), an easily administered multi-trial learning test (Lezak, 1995). In the VLT, fifteen monosyllabic words are presented in five subsequent trials, with a free recall procedure immediately following each presentation. After a delay of about 20 minutes, there is an additional free recall trial, followed by a recognition trial. Although the VLT is widely used in clinical practice and in memory research, its administration has not been uniform, which makes the finding of normative data difficult (Schmidt, 1996). Thus in this study, we sought to establish normative data for the VLT, using a procedure in which words were presented visually, on a computer screen, or auditorily, by means of an audio system. The normative data were derived from a cross-sectional study involving a large number of healthy participants ($N=1,855$) aged 24 to 81 years, matched for education and sex.

We decided to evaluate both auditory and visual stimulus presentation because earlier studies with other learning tasks suggested that presentation mode significantly affects the performance on short-term memory tasks. More specifically, auditorily presented material is recalled better than visually presented material. This phenomenon has been termed the ‘modality effect’ (see e.g. Greene, 1992; Penney, 1989), and has been observed in both younger and older adults (Taub, 1972, 1973). In contrast with short-term memory tasks, modality effects are conspicuously absent in long-term memory tasks (Penney, 1989). These results suggest that certain VLT measures, especially the Trial 1 score, might be significantly influenced by the mode of presentation, since it measures short-term memory. It is important to have normative data for two different sensory modality versions of the VLT (Van Boxtel et al., 2000) because subjects, and especially older individuals, may have visual or hearing impairments. This is especially relevant in the clinical assessment of older patients (La Rue, 1992), in whom a diagnosis of dementia may depend – in part – on the VLT performance score.

Previous research has suggested that the VLT performance is significantly affected by demographic variables, most notably age, sex, and education. For a comprehensive overview of previous VLT research, the reader is referred to Schmidt (1996). Geffen et al. (1990) found in a study of 153 healthy participants (age range 16- to 70+ years) that females outperformed males and that general VLT performance deteriorated with increasing age. Lannoo and Vingerhoets (1997) reported similar findings in their study of 200 healthy par-

ticipants aged 18 to 74 years. They found that females, young adults, and participants with a higher educational level performed better than males, older adults, and lower educated participants. Uchiyama and colleagues (1995) investigated the influence of test form, age, ethnicity, and education on VLT performance in a large sample of 2,056 healthy bisexual and homosexual HIV-seronegative males. Age was negatively and education positively related to general VLT performance. In contrast, Bolla-Wilson and Bleecker (1986) found that age affected performance on the first two learning trials only and that education did not have any effect in their study of 114 participants aged between 40 and 84 years. However, higher verbal intelligence was associated with a better performance as was sex, with females performing better than males. Savage and Gouvier (1992) investigated the influence of age and sex in a sample of 134 healthy participants aged 16- to 76 years. They reported a significant negative influence of increasing age on general VLT performance, whereas sex had no influence on test performance.

Since most – but not all – VLT studies suggested that demographic factors, such as age, sex and education, significantly affect VLT performance, the influence of these variables on VLT performance was investigated (next to presentation mode) in the present study in order to adequately stratify the normative data.

Method

Participants

The data used in this study were derived from the Maastricht Aging Study (Bosma et al., 2003; Jolles et al., 1995; Van Boxtel et al., 1998, 2000), a prospective study of determinants of cognitive aging. Participants were recruited from the Registration Network Family Practices, a sample frame for research in primary care (Metsemakers et al., 1992). Individuals were invited to participate in the study by their general practitioners, rather than by the project staff, because this was expected to improve participation and study compliance. All participants were aged between 24 and 81 years, and all were without documented medical conditions that could interfere with normal cognitive aging: individuals with chronic neurological pathology (e.g., dementia, cerebrovascular disease, epilepsy, parkinsonism, and malignancies related to the nervous system), mental retardation, psychopathology, or chronic psychotropic drug use were excluded. The sample was stratified for age (12 levels, ranging from 25±1 year, 30±1 year, ..., to 80±1 year), sex, and level of education (3 levels: low, average, and high). Level of education (LE) was assessed by classifying formal schooling in a system often used in the Netherlands (De Bie, 1987), which is comparable with the International Standard Classification of Education (UNESCO, Paris, 1976). Three groups were formed – those with at most primary education (LE low), those with junior vocational training (LE average),

and those with senior vocational or academic training (LE high). These three levels of education corresponded with an average of respectively 8.60, 11.41, and 15.25 years of full-time education in the sample (SD 1.95, 2.48, and 3.31, respectively).

Not all data for the 1,855 participants administered the VLT were included in the analyses. Data of participants who scored less than 24 on the Mini-Mental State Examination (Folstein et al., 1975) were excluded, since they could not be considered to be ‘cognitively intact’. For the auditory VLT version, an additional exclusion criterion was a hearing loss of 35 dB or more in the better ear. After these corrections – which excluded the data of 75 participants – the data of 1,780 participants were analysed (auditory VLT, $n=450$; visual VLT $n=1,330$). Basic descriptive data for the sample are provided in Table 1. The ethnic background of all participants was Caucasian, and all participants were native Dutch speakers.

Table 1. Descriptive characteristics of the sample ($N=1,780$).

Age Group	N	Age		Level of Education Frequency			Male:Female Ratio
		Mean	SD	Low	Average	High	
25±1 year	157	25.43	.90	21	75	61	78:79
30±1 year	151	30.62	.87	20	78	53	77:74
35±1 year	159	35.54	.91	34	76	49	79:80
40±1 year	156	40.51	.88	36	74	46	77:79
45±1 year	161	45.54	.92	50	71	40	80:81
50±1 year	161	50.30	.90	53	71	37	79:82
55±1 year	161	55.44	.90	84	50	27	82:79
60±1 year	158	60.60	.83	73	65	19	82:76
65±1 year	158	65.47	.87	86	52	20	78:80
70±1 year	153	70.29	.82	82	51	20	81:72
75±1 year	147	75.14	.78	78	47	21	69:78
80±1 year	58	79.83	.87	36	8	13	25:33
Total	1,780	51.29	16.22	653	718	406	887:893

Note. Data on level of education were missing from 3 participants.

Procedure and Instruments

All participants were tested individually at the neuropsychological laboratory of the Brain & Behaviour Institute (the Netherlands) using a Dutch adaptation of the VLT, the *Groningen Vijftienwoorden Test* (Groningen Fifteen Words Test), which had been developed earlier by Brand and Jolles (1985). These authors constructed six VLT versions using a pool of 410 monosyllabic words in the Dutch language. These words have a frequency of 20 to 400 per million according to the De La Court count of Dutch words (Linschoten, 1963). Only

common words were used which are acquired early in life (before age 6) and refer to concrete objects such as flower, crown, bear, tree, etcetera. Potentially ambiguous words and words referring to emotional factors were not used. Various possible categories of words (e.g. animals, body parts, parts of the house, furniture) were evenly distributed over the lists in order to control for potential semantic associations. Auditory associations were controlled for per list. The test versions were evaluated in control experiments after which particular words were discarded and replaced by new words. The test versions were eventually shown to be parallel and used in numerous drug studies in which repeated administration of parallel versions of tests is essential (e.g. Hijman et al., 1992). In the present study, we used one of these six parallel versions.

The test instructions were read aloud by the experimenter. The participant was required to listen to, or to watch, the fifteen words that were presented by means of an audio system (auditory version) or on a computer screen (visual version) and to repeat as many words as possible as soon as the presentation stopped. There was no restriction on the output order. The first trial was followed by four more trials in which the words were presented in identical order. When the fifth trial was completed, a fixed battery of other cognitive tests was administered for about 20 minutes. These tests did not involve the learning of verbal material in order to avoid interference with the previously learned VLT words (for a description of these tests, see Jolles et al. (1995)). After the delay – and unexpectedly for the participants – the instruction was given to recall the words learned (delayed recall). This was followed immediately by a recognition test, involving yes/no recognition of the fifteen words intermixed with fifteen non-target words.

Stimulus words were presented for 1 second, with an inter-stimulus interval of 1 second. In the auditory VLT, the words were presented by the digitized voice of a professional speaker, using a high-fidelity stereo speaker system connected to a standard PC-microcomputer. In the visual VLT, stimuli were presented on a 17-inch computer screen operated by a standard PC-microcomputer. All words were displayed in capital letters of the font 'Times' size 72, in white against a black background and in the centre of the screen. If necessary, the visual VLT was performed with optical correction for nearby vision.

The auditory VLT version was administered to 1/4th of the sample, and the visual VLT to 3/4th of the sample. The participants were divided at random over the conditions on a 1:3 basis (ratio of the number of participants in the auditory versus visual VLT condition) and with the restriction that the proportion of participants that were assessed with the visual versus auditory VLT was equal for each Age group x Level of education x Sex cell.

The data were collected by five test assistants who had been intensively trained in test administration by the neuropsychologists and physician of the project staff. Members of the project staff visited the test assistants during assessment of the VLT at least once a week to ensure uniform test administration and data collection.

Outcome measures

The following measures were analysed: (i) Trial 1, 2, 3, 4 and 5 score, the number of correctly recalled words in each of these learning trials. (ii) Repetitions, i.e. the total number of words repeated over Trials 1 to 5. Only words that were repeated during the first recall-attempt per trial were scored as repetitions (e.g. flower – crown – bear – crown – ...). This precludes that words repeated during the often-seen strategy of overtly repeating a sequence of words are scored as repetitions. (iii) Delayed recall, the number of correctly recalled words after the 20-minute delay. (iv) Total recall 1-5, the total number of correctly recalled words summed over the five learning trials. (v) Total recall 1-3, the total number of correctly recalled words summed over the first three learning trials. (vi) Delta, which was calculated as the Trial 3 score minus the Trial 1 score. The latter two scores were used because ceiling effects are often observed in the VLT, especially in younger adults: the increase in the number of words recalled is especially pronounced during the first three learning trials, which makes the scores for the last two learning trials to some extent redundant. Delta differs from the Total recall 1-3 score because it is a learning measure that corrects for the Trial 1 score, which is a measure of short-term and working memory. (vii) Recognition, calculated as the number of correct hits minus the number of false positives. This measure is equivalent to the Recognition Percent Correct score ($= [\text{true positives} + \text{true negatives}] / 30$) proposed by Harris et al. (2002). (viii) A Retention score, calculated as the Delayed recall score divided by the Trial 5 score. Thus a retention score of 0.75 indicates that the individual's score on the delayed recall score was 25% lower than that person's score in Trial 5.

Data analysis

Multiple linear regression models were fitted using a step-down hierarchical procedure for the VLT measures adjusted for Age, Age² (which enables modelling of quadratic age effects), Sex, Level of education, and Mode of presentation. Since previous VLT research provided equivocal results with respect to the influence of demographic factors on VLT performance – especially with respect to education and sex – the interactions between Age and Education, Age and Sex, Age² and Education, and Age² and Sex were also included in the model. Age was centered (Age = calendar age minus 50) before the quadratic terms and interactions were computed. This was done in order to avoid multicollinearity (Neter et al., 1989), because high levels of multicollinearity may lead to technical problems when estimating the regression coefficients (Aiken & West, 1991). Level of education (LE) was dummy coded with LE average as the reference category. The dummies LE low and LE high were always either both included in, or both excluded from, the model, since they belong together and represent

the effect of the categorical predictor education. Similarly, their interactions with another predictor were always either both included in or excluded from the model. Non-significant predictors ($p > .005$; A lower *alpha* level was chosen in order to avoid Type I errors due to multiple testing) were excluded from the model, but no predictor was removed from the model as long as it was also included in a higher order term in the model. In particular, Age was never removed if Age² or any interaction involving Age or Age² was still in the model. The reason for this is that the p -value of any predictor is arbitrary (depending on the coding used for predictors) if that predictor is part of a higher-order predictor in the model (Aiken & West, 1991). The assumptions of regression analysis (homoscedasticity, normal distribution of the residuals, absence of multicollinearity, and absence of 'influential cases') were tested for each model. Homoscedasticity was evaluated by visual inspection of the scatter plots of the residuals against the predicted values. Normal distribution of the residuals was investigated by visual inspection of the histograms and the normal probability plots. The occurrence of multicollinearity was checked by calculating the Variance Inflation Factors (VIFs), which should not exceed 10 (Belsley et al., 1980). Cook's distances were calculated to identify possible influential cases. All analyses were performed using the SPSS 10.0 for Macintosh software package.

Normative data can then be obtained by calculating the residuals for each VLT score ($e_i = \text{observed score} - \text{predicted score}$): after standardization of the residuals ($Z_i = e_i / \text{SD}[\text{residual}]$), the performance of the participant can be evaluated via a Z distribution table with cumulative probabilities.

Results

The regression models are presented in Table 2. No significant influence of outliers was observed (maximum Cook's distance .07). The VIFs of the predictors in the regression models were at most 2.7, well below the cut-off value. The Repetitions score was square root transformed ($\sqrt{\text{Repetitions}}$) because of positively skewed residuals before transformation of the score.

Age was by far the strongest predictor of performance on all VLT scores (see standardized B's, Table 2), with the exception of Delta. An additional quadratic effect of age on VLT performance was observed for most VLT scores, with the exception of Delta, although the quadratic trend was less important than the linear age effect. Females and highly educated participants scored significantly higher than males and their less educated counterparts on most VLT measures. Mode of presentation differentially affected VLT performance: the Trial 1 score was higher with the auditory VLT compared to the visual VLT, and the opposite was found for the Trial 3, 4, Delta and $\sqrt{\text{Repetitions}}$ score. The influence of presentation mode on the VLT Trials 1 to 5 and the Delayed recall trial performance is shown in Figure 1, for four broad age groups (24-36 years, 39-51 years, 54-66 years, and 69-81 years).

Table 2. Multiple linear regression models of the VLT scores following a step-down hierarchical procedure; the full model included Age, Age², LE low, LE high, Sex, Mode of Presentation, Age x LE low, Age x LE high, Age x Sex, Age² x LE low, Age² x LE high and Age² x Sex.

Score	Variable	B	Std. Error B	T	Standardized B	R ²
Trial 1	(constant)	5.627	.085	66.323**		
	Age	-.048	.004	-12.644**	-.422	
	Age ²	-.0005	0	-2.959*	-.064	
	Sex	-.425	.076	-5.578**	-.116	
	LE low	-.537	.094	-5.745**	-.141	
	LE high	.417	.102	4.092**	.095	
	Mode of presentation	.353	.087	4.059**	.083	
	Age x LE low	.018	.006	3.057*	.095	
	Age x LE high	-.0074	.006	-1.187	-.032	.259
Trial 2	(constant)	8.972	.098	91.097**		
	Age	-.052	.003	-17.113**	-.378	
	Age ²	-.0006	0	-2.899*	-.060	
	Sex	-.808	.092	-8.751**	-.181	
	LE low	-.594	.108	-5.483**	-.128	
	LE high	.514	.121	4.261**	.097	.257
Trial 3	(constant)	10.909	.109	100.340**		
	Age	-.053	.003	-16.510**	-.363	
	Age ²	-.0008	0	-3.771**	-.078	
	Sex	-.971	.098	-9.904**	-.204	
	LE low	-.678	.115	-5.890**	-.137	
	LE high	.491	.128	3.833**	.087	
	Mode of presentation	-.493	.112	-4.408**	-.090	.265
Trial 4	(constant)	11.865	.107	111.229**		
	Age	-.050	.003	-15.915**	-.349	
	Age ²	-.0009	0	-4.776**	-.099	
	Sex	-.995	.096	-10.343**	-.213	
	LE low	-.691	.113	-6.121**	-.142	
	LE high	.523	.126	4.157**	.094	
	Mode of presentation	-.358	.110	-3.266**	-.066	.265
Trial 5	(constant)	12.462	.103	121.489**		
	Age	-.046	.003	-14.632**	-.327	
	Age ²	-.0007	0	-3.640**	-.077	
	Sex	-1.018	.096	-10.584**	-.222	
	LE low	-.670	.113	-5.932**	-.141	
	LE high	.497	.126	3.956**	.091	.237

Total recall 1-5	(constant)	49.672	.401	123.945**		
	Age	-.247	.012	-19.916**	-.416	
	Age ²	-.0033	.001	-4.250**	-.084	
	Sex	-4.227	.376	-11.253**	-.220	
	LE low	-3.055	.441	-6.925**	-.153	
	LE high	2.496	.491	5.081**	.109	.334
Total recall 1-3	(constant)	25.440	.243	104.816**		
	Age	-.150	.007	-19.968**	-.422	
	Age ²	-.0016	0	-3.454**	-.069	
	Sex	-2.217	.227	-9.748**	-.193	
	LE low	-1.699	.267	-6.361**	-.143	
	LE high	1.467	.297	4.932**	.107	.318
Delta	(constant)	5.100	.068	74.881**		
	Age	-.013	.003	-4.747**	-.109	
	Sex	-.502	.089	-5.651**	-.130	
	Mode of presentation	-.858	.102	-8.385**	-.192	.064
Delayed recall	(constant)	10.924	.128	85.374**		
	Age	-.073	.004	-18.440**	-.401	
	Age ²	-.0009	0	-3.704**	-.076	
	Sex	-1.197	.120	-9.985**	-.203	
	LE low	-.844	.141	-5.991**	-.138	
	LE high	.424	.157	2.707	.060	.282
Retention	(constant)	.871	.007	125.526**		
	Age	-.003	0	-13.652**	-.310	
	Age ²	-.00005	0	-3.405*	-.077	
	Sex	-.029	.008	-3.812**	-.085	.116
Recognition	(constant)	14.191	.090	157.143**		
	Age	-.034	.003	-10.757**	-.248	
	Age ²	-.0008	0	-3.867**	-.089	
	Sex	-.398	.101	-3.957**	-.090	.084
√Repetitions	(constant)	2.052	.040	50.947**		
	Age	-.013	.001	-9.815**	-.226	
	Age ²	-.0004	0	-3.945**	-.091	
	Sex	-.129	.043	-3.021*	-.068	
	Mode of presentation	-.280	.049	-5.702**	-.129	.086

Note. LE = Level of education. Coding of the predictors: Age = calendar age - 50; Age² = (calendar age - 50)²; Sex: Male = 1, Female = 0; LE Low: Low Education = 1, Average or High Education = 0; LE High: High Education = 1, Low or Average Education = 0; Mode of Presentation: Auditory version = 1, Visual version = 0.
 * $p < .005$; ** $p < .001$

Figure 1. Mean VLT Trial 1, 2, 3, 4, 5 and Delayed recall scores for four Age groups (24-36 years, 39-51 years, 54-66 years and 69-81 years), separated for Presentation mode.

Only one significant interaction was found: an Age x LE low interaction on the Trial 1 score. Figure 2 depicts the predicted Trial 1 scores for male participants for the visual VLT (the shape of these plots was identical for females and for the auditory VLT, with a constant value added or subtracted to the predicted scores – see Table 2). As shown in Figure 2, low educated participants performed worse compared to average/high educated participants for all ages, but the difference in performance between the level of education groups decreased with increasing age.

Figure 2. Predicted VLT Trial 1 scores for male participants on the visual VLT, per Level of education category as a function of Age.

Normative data are obtained by calculating the participant's predicted VLT scores, using the regression models (see Table 2). The residuals of each score are then calculated (e_i = observed score - predicted score), and standardized ($Z_i = e_i / \text{SD}[\text{residual}]$) using Table 3. Note that the scatter plots of the residuals against the predicted values showed a trend to heteroscedasticity (a decrease in the variance as a function of the predicted scores) for the Recognition and Retention scores. Therefore, the standard deviations of the residuals of these two scores were calculated as a function of the predicted scores divided into quartiles (Table 3). After standardization of the residuals, the participant's performance can be evaluated via a Z distribution table with cumulative probabilities (or a simplified version thereof, see Table 4).

Table 3. Standard deviations of the residuals per VLT measure.

Measure	Predicted score	SD(residual)
Trial 1	General (all scores)	1.584
Trial 2	General (all scores)	1.923
Trial 3	General (all scores)	2.042
Trial 4	General (all scores)	2.003
Trial 5	General (all scores)	2.003
Total recall 1-5	General (all scores)	7.826
Total recall 1-3	General (all scores)	4.739
Delta	General (all scores)	1.873
Delayed recall	General (all scores)	2.496
Retention	General (all scores)	0.163
	≤ 0.795	0.191
	Between 0.796 and 0.852	0.172
	Between 0.853 and 0.886	0.154
	≥ 0.887	0.127
	General (all scores)	2.119
	≤ 13.285	2.972
	Between 13.286 and 13.937	2.250
	Between 13.938 and 14.172	1.549
	≥ 14.172	1.300
√ Repetitions	General (all scores)	0.899

For example, the visual VLT was administered to a 60-year-old man with a low level of education. The participant recalled 5 words in the Delayed recall trial. The predicted Delayed recall score for this person would be 8.063 (= $10.924 + ((60 - 50) * -.073) + ((60 - 50)^2 * -.0009) + (1 * -.197) + (1 * -.844) + (0 * .424)$), with a residual of - 3.063 ($e_i = 5 - 8.063$). The SD(residual) equals

2.496 and the standardized residual -1.227 ($= -3.063 / 2.496$), which corresponds to a p -value of .11, and can be considered “below normal” according to Table 4.

Table 4. Z score intervals with the corresponding proportion of individuals with a score falling within this interval and corresponding verbal labels.

Z Score interval	Percentage within this interval	Verbal label
Below -2.0	2.3 %	Very poor performance
Between -2.0 and -1.5	4.4 %	Poor performance
Between -1.5 and -1.0	9.2 %	Below normal
Between -1.0 and -0.5	15 %	Normal
Between -0.5 and 0.5	38.3 %	Normal
Between 0.5 and 1.0	15 %	Normal
Between 1.0 and 1.5	9.2 %	Above normal
Between 1.5 and 2.0	4.4 %	Good performance
Above 2.0	2.3 %	Very good performance

Tables A1-A4 (in Appendix) provide normative tables based on the regression models in Table 2 for the Trial 1, Delta, Delayed recall, and Recognition scores, stratified by their statistically significant predictors. Due to space constraints, we only provide normative tables for the most important VLT measures here. Other scores can be evaluated using the described normative procedure. In the course of 2005, a normative handbook will also be made available via Neuropsych Publishers (<http://www-np.unimaas.nl/>).

If an individual is not exactly 25, 30, ..., 80 years old, then the person's age should be rounded up to the closest age given. If the performance of such an individual is borderline-normal according to these tables (e.g. Z value ≈ -1.28), the regression models presented in Table 2 can be used to determine the exact Z values.

Discussion

The results of the present study showed highly significant effects of age on VLT performance. Interestingly, the age-related decline in performance started early in life, in a highly linear fashion: the linear age-effect was a much stronger predictor of VLT performance than the quadratic age-effect. Put in another way, the age-related differences in general VLT performance were more manifest in a quantitative than in a qualitative way – although the absolute number of recalled words differed as a function of age, the shape of the learning curve was similar (see Figure 1). The results suggested that younger adults outperformed older adults because of a better performance on Trial 1. These differences in Trial 1 performance may be due to differences in processing speed, as suggested

by Bolla-Wilson and Bleecker (1986). In clinical practice, the Total recall 1-5 score is the most commonly used learning summary score (Schmidt, 1996). The results of the present study, however, suggest that this measure is not the most appropriate one. Such a summary measure suggests that the differences in learning capacity over the first five trials in the VLT vary largely as a function of age, which was not supported by the results of the present study (see Figure 1). The same criticism applies to the Total recall 1-3 score, although this score has the advantage over the Total recall 1-5 score in that it takes ceiling effects, at least theoretically, into account. A measure such as Delta (= Trial 3 - Trial 1 score) might be a more appropriate learning measure, since it corrects for Trial 1 performance and takes ceiling effects into account, and thus gives a more realistic estimation of the verbal learning capacity of an individual.

Although Schmidt (1996) suggested that the effect of sex appears to be of limited importance in interpreting VLT results, the present study clearly showed significant effects of sex on all VLT measures, with women consistently outperforming men. This finding is in agreement with studies that point to the superior verbal abilities of women, for example on verbal episodic memory tasks (e.g. Herlitz & Yonker, 2002; Yonker et al., 2003) and verbal fluency tasks (e.g. Rideout & Winchester, 1990). With regard to education, the negative impact of a low level of education - as compared to an average level of education - on VLT performance, was larger than the positive impact of a high level of education on most VLT scores (see the standardized B values in Table 2). Furthermore, the significant Age x Low level of education interaction on the Trial 1 score suggested that the differences in performance between the low level of education group and the average/high level of education group decreased with increasing age.

It should be noted that Bolla-Wilson and Bleecker (1986) found no effect of education on VLT performance, in contrast to the results of the present study. This difference is most likely attributable to the fact that Bolla-Wilson and Bleecker (1986) used education together with a verbal intelligence measure as predictors of the VLT performance (next to age and sex). It is possible that a variable such as (verbal) intelligence is a more appropriate estimator of a person's general ability than is education, especially in older individuals. Indeed, before the 1960s most people did not continue their schooling, mostly for reasons other than their intellectual capacity, e.g. for socio-economic reasons (Jolles et al., 1995). Nevertheless, we used level of education as an estimator of general ability instead of verbal intelligence for two main reasons. First, in order to facilitate the use of normative data, it is important to use a readily available stratification variable. Use of verbal intelligence as a predictor of the VLT performance means that a person would also have to complete a verbal intelligence test before the normative VLT data could be used. This is time-consuming and expensive, and might be difficult in certain groups, e.g. elderly people, people with attention disorders, people with a hearing impairment, etc. Second, as with any measurement of a psychological function, there is an error component in the

obtained score of a verbal intelligence test. This would introduce an extra source of error to the model (the error component in the score on the verbal intelligence test) in addition to the error component for the different VLT measures.

Mode of stimulus presentation had a significant, but not uniform, influence on some VLT measures. Thus the Trial 1 score was higher with the auditory VLT (in accordance with the modality effect), the Trial 3, 4, Delta and $\sqrt{\text{Total}}$ repetition scores were higher with the visual VLT. It should be noted that in the models predicting the Delayed recall and Trial 5 scores, the predictor Mode of presentation was deleted in the final step of the hierarchical procedure, but was almost significant ($p = .008$ and $p = .010$, respectively). Thus, the pattern of results suggests that auditory presentation is associated with a better performance of tasks depending mainly on short-term or working memory processes (VLT Trial 1), whereas visual presentation is associated with a better performance of tasks that depend on long-term memory processes (VLT Delayed recall trial) or tasks that depend on a mixture of short-term and long-term memory processes (VLT Trials 3, 4, and 5). This stresses the importance of using normative data appropriate for the test procedure (stimulus presentation) used. This aspect tends to be overlooked in clinical practice, when tests are computerized even though the normative data were obtained for non-computerized tests.

At present, it is not known whether the visual VLT or the auditory VLT version should be used as first choice to assess verbal memory or to detect pathological conditions such as e.g. dementia. This issue merits further investigation. However, an important issue to keep in mind regarding this question concerns earlier findings by Van Boxtel and colleagues (2000). These authors suggested that the auditory VLT version underestimates the true memory capacity of people with mild to moderate degrees of hearing loss, even after the volume at which the VLT words are presented is adjusted for individual hearing ability. Importantly, Van Boxtel et al. (2000) also reported a large overlap in hearing acuity between people who did or did not report hearing loss. Therefore, we recommend that the visual VLT version be used to test participants with an unknown (objective) hearing status (even if the participant claims not to have a hearing loss) or with a (even mild) hearing impairment.

The present data were obtained for a healthy population of native Dutch speakers. It can be questioned whether the normative data of the Dutch VLT are applicable to individuals who have a different native language or cultural background. Recent studies suggested that words that are selected using the same criteria (e.g. comparable with regard to semantic and phonemic features, words that seem equally concrete, words that seem as difficult or easy to cluster by association, etc.) lead to comparable test versions across countries and languages (Van den Burg & Kingma, 1999). This was the case when we compared our data for the auditory VLT version with the results of Geffen and colleagues (1990) obtained with an English-speaking sample. After re-defining the age categories of our sample to match those of Geffen et al. (1990) (see Figure 3),

we found that the mean VLT Trial 1, 2, 3, 4, 5, and Total recall 1-5 scores in our sample were not significantly different from the corresponding mean scores in their sample (all $p > .01$) for all age by sex groups (see Figure 3 for the Total recall 1-5 score). Note that a series of two-tailed t -tests per age by sex group were performed instead of a more appropriate technique, for example ANOVA with Age category and Sex as between subject factors, since we did not have access to the raw dataset of Geffen et al. (1990).

Figure 3. Mean auditory VLT Total recall 1-5 scores for males (left) and females (right) separated for Age category (and 95% Confidence Interval).

We were not able to compare the scores for the visual VLT (because Geffen and colleagues did not administer a visual VLT version), and for certain VLT measures, such as Recognition and Delayed recall (because Geffen and colleagues used different test administration and scoring procedures, which means that these scores are not comparable). However, we expect it to be unlikely that English-speaking people would score differently from Dutch-speaking people on a visual version of the VLT, especially when differences were not found with the auditory version of the VLT. Supporting evidence for this claim can also be found in a study of Møller and colleagues (1998), who showed that the Visual Verbal Learning Test (in three trials instead of five) was culturally robust when evaluated in seven different language areas (English, Danish, Dutch, French, German, Greek, and Spanish). Thus the results of these studies suggest that normative data for a Dutch sample (and normative data for a sample with a language very similar to Dutch, such as German; see Heubrock, 1994) can also be used for samples of English-speaking people.

In summary, we established normative data for the auditory and visual version of the VLT. The data were derived from a large sample of 1,855 cognitively intact men and women of different educational levels and ages (range 24-

81 years). We used multiple regression models to establish normative data that account for the effects of these variables. The results showed that VLT performance is superior in women, decreases with age, and is associated with level of education. Performance on VLT Trial 1, which depends mainly on short-term or working memory, was better with the auditory version, whereas performance on VLT trials that depend on long-term memory processes or trials that depend on a combination of short-term and long-term memory processes was better with the visual version. Furthermore, although the Total recall 1-5 score is currently the most frequently used VLT learning summary measure, our results indicated that Delta ($= \text{Trial 3} - \text{Trial 1}$) may provide a better estimation of the verbal learning capacity of an individual.

Appendix

Table A1. Normative VLT data for the Trial 1 score stratified by Mode of presentation, Level of education, Sex, and, Age (25, 30, ..., 80 years). The raw test score leading to a particular Z-value is given for Z-values indicating the percentiles 5, 10, 20, 50, 80, 90, and 95.

	VISUAL VLT																										
	LE low	Z	Cum. prob.	MALE													FEMALE										
				Age in years													Age in years										
				25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
		1.64	.95	7.7	7.7	7.6	7.5	7.4	7.3	7.1	6.9	6.7	6.5	6.2	5.9	8.1	8.1	8.0	7.9	7.8	7.7	7.5	7.3	7.1	6.9	6.6	6.3
		1.28	.90	7.1	7.1	7.0	6.8	6.7	6.5	6.3	6.1	5.9	5.6	5.3	7.6	7.6	7.5	7.5	7.4	7.3	7.1	7.0	6.8	6.6	6.3	6.1	5.8
		0.84	.80	6.4	6.4	6.3	6.2	6.1	6.0	5.8	5.6	5.4	5.2	4.9	4.6	6.9	6.8	6.8	6.7	6.6	6.4	6.3	6.1	5.9	5.6	5.4	5.1
		0	.50	5.1	5.1	5.0	4.9	4.8	4.7	4.5	4.3	4.1	3.9	3.6	3.3	5.5	5.5	5.4	5.3	5.2	5.1	4.9	4.7	4.5	4.3	4.0	3.7
		-0.84	.20	3.8	3.7	3.7	3.6	3.5	3.3	3.2	3.0	2.8	2.5	2.3	2.0	4.2	4.2	4.1	4.0	3.9	3.8	3.6	3.4	3.2	3.0	2.7	2.4
		-1.28	.10	3.1	3.0	3.0	2.9	2.8	2.6	2.5	2.3	2.1	1.8	1.6	1.3	3.5	3.5	3.4	3.3	3.2	3.1	2.9	2.7	2.5	2.3	2.0	1.7
	-1.64	.05	2.5	2.5	2.4	2.3	2.2	2.1	1.9	1.7	1.5	1.3	1.0	0.7	2.9	2.9	2.8	2.7	2.6	2.5	2.3	2.1	1.9	1.7	1.4	1.1	
		Z	Cum. prob.	25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
		1.64	.95	8.7	8.6	8.4	8.2	8.0	7.8	7.5	7.3	7.0	6.6	6.3	5.9	9.1	9.0	8.8	8.7	8.5	8.2	8.0	7.7	7.4	7.1	6.7	6.3
		1.28	.90	8.1	8.0	7.8	7.7	7.5	7.2	7.0	6.7	6.4	6.1	5.7	5.3	8.5	8.4	8.3	8.1	7.9	7.7	7.4	7.1	6.8	6.5	6.1	5.8
		0.84	.80	7.4	7.3	7.1	7.0	6.8	6.5	6.3	6.0	5.7	5.4	5.0	4.6	7.8	7.7	7.6	7.4	7.2	7.0	6.7	6.4	6.1	5.8	5.4	5.1
		0	.50	6.1	6.0	5.8	5.6	5.4	5.2	4.9	4.7	4.4	4.0	3.7	3.3	6.5	6.4	6.2	6.1	5.9	5.6	5.4	5.1	4.8	4.5	4.1	3.7
		-0.84	.20	4.8	4.6	4.5	4.3	4.1	3.9	3.6	3.3	3.0	2.7	2.4	2.0	5.2	5.1	4.9	4.7	4.5	4.3	4.0	3.8	3.5	3.1	2.8	2.4
		-1.28	.10	4.1	3.9	3.8	3.6	3.4	3.2	2.9	2.6	2.3	2.0	1.7	1.3	4.5	4.4	4.2	4.0	3.8	3.6	3.3	3.1	2.8	2.4	2.1	1.7
	-1.64	.05	3.5	3.4	3.2	3.0	2.8	2.6	2.4	2.1	1.8	1.4	1.1	0.7	3.9	3.8	3.6	3.5	3.3	3.0	2.8	2.5	2.2	1.9	1.5	1.1	
		Z	Cum. prob.	25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
		1.64	.95	9.3	9.1	8.9	8.7	8.5	8.2	7.9	7.6	7.3	6.9	6.5	6.1	9.7	9.5	9.4	9.1	8.9	8.6	8.4	8.0	7.7	7.3	6.9	6.5
		1.28	.90	8.7	8.6	8.4	8.2	7.9	7.6	7.4	7.0	6.7	6.3	5.9	5.5	9.1	9.0	8.8	8.6	8.3	8.1	7.8	7.5	7.1	6.8	6.4	6.0
		0.84	.80	8.0	7.9	7.7	7.5	7.2	6.9	6.7	6.3	6.0	5.6	5.3	4.8	8.4	8.3	8.1	7.9	7.6	7.4	7.1	6.8	6.4	6.1	5.7	5.3
		0	.50	6.7	6.5	6.3	6.1	5.9	5.6	5.3	5.0	4.7	4.3	3.9	3.5	7.1	7.0	6.8	6.5	6.3	6.0	5.8	5.4	5.1	4.7	4.3	3.9
		-0.84	.20	5.4	5.2	5.0	4.8	4.6	4.3	4.0	3.7	3.3	3.0	2.6	2.2	5.8	5.6	5.4	5.2	5.0	4.7	4.4	4.1	3.8	3.4	3.0	2.6
		-1.28	.10	4.7	4.5	4.3	4.1	3.9	3.6	3.3	3.0	2.6	2.3	1.9	1.5	5.1	4.9	4.7	4.5	4.3	4.0	3.7	3.4	3.1	2.7	2.3	1.9
	-1.64	.05	4.1	3.9	3.7	3.5	3.3	3.0	2.7	2.4	2.1	1.7	1.3	0.9	4.5	4.4	4.2	4.0	3.7	3.4	3.2	2.8	2.5	2.1	1.7	1.3	

(Continued on the next page)

AUDITORY VLT	LE low	Z	Cum. prob.	MALE													FEMALE												
				Age in years													Age in years												
				25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80		
LE average		1.64	.95	8.1	8.0	7.9	7.8	7.6	7.5	7.3	7.1	6.8	6.6	6.3	8.5	8.4	8.4	8.3	8.2	8.0	7.9	7.7	7.5	7.2	7.0	6.7			
		1.28	.90	7.5	7.4	7.4	7.3	7.2	7.0	6.9	6.7	6.5	6.2	6.0	5.7	7.9	7.9	7.8	7.7	7.6	7.5	7.3	7.1	6.9	6.7	6.4			
		0.84	.80	6.8	6.7	6.7	6.6	6.5	6.3	6.2	6.0	5.8	5.5	5.3	5.0	7.2	7.2	7.1	7.0	6.9	6.8	6.6	6.4	6.2	6.0	5.7			
		0	.50	5.5	5.4	5.4	5.3	5.2	5.0	4.9	4.7	4.5	4.2	4.0	3.7	5.9	5.8	5.8	5.7	5.6	5.4	5.3	5.1	4.9	4.6	4.4			
		-0.84	.20	4.1	4.1	4.0	3.9	3.8	3.7	3.5	3.3	3.1	2.9	2.6	2.3	4.5	4.5	4.4	4.4	4.2	4.1	3.9	3.8	3.5	3.3	3.0			
		-1.28	.10	3.4	3.4	3.3	3.2	3.1	3.0	2.8	2.6	2.4	2.2	1.9	1.6	3.9	3.8	3.8	3.7	3.6	3.4	3.3	3.1	2.9	2.6	2.4			
		-1.64	.05	2.9	2.8	2.8	2.7	2.6	2.4	2.3	2.1	1.9	1.6	1.4	1.1	3.3	3.2	3.2	3.1	3.0	2.8	2.7	2.5	2.3	2.0	1.8			
LE high		Z	Cum. prob.	Age in years													Age in years												
		1.64	.95	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0		
		1.28	.90	8.5	8.3	8.2	8.0	7.8	7.6	7.3	7.1	6.8	6.4	6.1	5.7	8.9	8.8	8.6	8.4	8.2	8.0	7.8	7.5	7.2	6.8	6.5	6.1		
		0.84	.80	7.8	7.6	7.5	7.3	7.1	6.9	6.6	6.4	6.1	5.7	5.4	5.0	8.2	8.1	7.9	7.7	7.5	7.3	7.1	6.8	6.5	6.2	5.8	5.4		
		0	.50	6.4	6.3	6.2	6.0	5.8	5.6	5.3	5.0	4.7	4.4	4.0	3.7	6.9	6.7	6.6	6.4	6.2	6.0	5.7	5.5	5.1	4.8	4.5	4.1		
		-0.84	.20	5.1	5.0	4.8	4.7	4.5	4.2	4.0	3.7	3.4	3.1	2.7	2.3	5.5	5.4	5.3	5.1	4.9	4.6	4.4	4.1	3.8	3.5	3.1	2.8		
		-1.28	.10	4.4	4.3	4.1	4.0	3.8	3.5	3.3	3.0	2.7	2.4	2.0	1.6	4.8	4.7	4.6	4.4	4.2	4.0	3.7	3.4	3.1	2.8	2.4			
		-1.64	.05	3.8	3.7	3.6	3.4	3.2	3.0	2.7	2.4	2.1	1.8	1.4	1.1	4.3	4.1	4.0	3.8	3.6	3.4	3.1	2.9	2.5	2.2	1.9	1.5		

Note. LE = Level of education; Cum. prob. = cumulative probability.

Table A2. Normative VLT data for the Delta score stratified by Mode of presentation, Sex, and, Age (25, 30, ..., 80 years). The raw test score leading to a particular Z-value is given for Z-values indicating the percentiles 5, 10, 20, 50, 80, 90 and 95.

VISUAL VLT	Z value	Cum. prob.	MALE												FEMALE												
			Age in years												Age in years												
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80	
LE low/average/high	1.64	.95	8.0	7.9	7.9	7.8	7.7	7.7	7.6	7.5	7.4	7.3	7.3	8.5	8.4	8.4	8.3	8.2	8.2	8.1	8.0	8.0	7.9	7.8	7.8		
	1.28	.90	7.3	7.3	7.2	7.1	7.1	7.0	6.9	6.9	6.8	6.7	6.6	7.8	7.8	7.7	7.6	7.6	7.5	7.4	7.4	7.3	7.2	7.2	7.1		
	0.84	.80	6.5	6.4	6.4	6.3	6.2	6.2	6.1	6.0	6.0	5.9	5.8	7.0	6.9	6.9	6.8	6.7	6.7	6.6	6.5	6.5	6.4	6.3	6.3		
	0	.50	4.9	4.9	4.8	4.7	4.7	4.6	4.5	4.5	4.4	4.3	4.2	5.4	5.4	5.3	5.2	5.2	5.1	5.0	5.0	4.9	4.8	4.7	4.7		
	-0.84	.20	3.3	3.3	3.2	3.2	3.1	3.0	3.0	2.9	2.8	2.8	2.7	2.6	3.9	3.8	3.7	3.7	3.6	3.5	3.5	3.4	3.3	3.3	3.2		
	-1.28	.10	2.5	2.5	2.4	2.3	2.3	2.2	2.1	2.1	2.0	1.9	1.9	1.8	3.0	3.0	2.9	2.8	2.8	2.7	2.6	2.6	2.5	2.4	2.3	2.3	
AUDITORY VLT	-1.64	.05	1.9	1.8	1.7	1.7	1.6	1.5	1.5	1.4	1.3	1.3	1.2	1.1	2.4	2.3	2.2	2.2	2.1	2.0	1.9	1.8	1.7	1.6	1.6		
	LE low/average/high	Cum. prob.	MALE												FEMALE												
			Age in years												Age in years												
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80	
			1.64	.95	7.1	7.1	7.0	6.9	6.8	6.7	6.7	6.6	6.6	6.5	6.4	7.6	7.6	7.5	7.4	7.4	7.3	7.2	7.2	7.1	7.1	7.0	6.9
			1.28	.90	6.5	6.4	6.3	6.2	6.1	6.1	6.0	5.9	5.9	5.8	5.7	7.0	6.9	6.8	6.8	6.7	6.6	6.6	6.5	6.4	6.4	6.3	6.2
			0.84	.80	5.6	5.6	5.5	5.4	5.4	5.3	5.2	5.2	5.1	5.0	4.9	6.1	6.1	6.0	5.9	5.9	5.8	5.8	5.7	5.6	5.6	5.5	5.4
LE low/average/high	0	.50	4.1	4.0	3.9	3.9	3.8	3.7	3.7	3.6	3.5	3.5	3.4	4.6	4.5	4.4	4.4	4.3	4.2	4.2	4.1	4.0	4.0	3.9	3.9		
	-0.84	.20	2.5	2.4	2.4	2.3	2.2	2.2	2.1	2.0	2.0	1.9	1.8	3.0	2.9	2.9	2.8	2.7	2.7	2.6	2.5	2.5	2.4	2.3	2.3		
	-1.28	.10	1.7	1.6	1.5	1.5	1.4	1.3	1.3	1.2	1.1	1.1	1.0	2.2	2.1	2.0	2.0	1.9	1.8	1.8	1.7	1.6	1.6	1.5	1.5		
	1.0	0.9	0.9	0.8	0.7	0.7	0.6	0.5	0.5	0.4	0.3	0.3	1.5	1.4	1.4	1.3	1.2	1.2	1.1	1.0	1.0	0.9	0.8	0.8	0.8		

Note. LE = Level of education; Cum. prob. = cumulative probability.

Table A3. Normative VLT data for the Recognition score stratified by Sex and Age (25, 30, ..., 80 years). The raw test score leading to a particular Z-value is given for Z-values indicating the percentiles 5, 10, 20, 50, 80, 90 and, 95.

AUDITORY & VISUAL VLT		Z value		Cum. prob.	MALE										FEMALE											
					Age in years										Age in years											
					25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70
LE low/average/high	1.64	.95	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0
	1.28	.90	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0
	0.84	.80	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	14.9	14.5	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	15.0	14.9	14.5
	0	.50	14.1	14.2	14.1	14.1	13.9	13.8	13.6	13.4	13.1	12.8	12.4	12.1	14.5	14.6	14.5	14.5	14.3	14.2	14.0	13.8	13.5	13.2	12.8	12.5
	-0.84	.20	12.8	12.9	12.8	12.8	12.6	11.9	11.7	11.5	10.6	10.3	9.9	9.6	13.4	13.5	13.4	13.4	13.2	13.1	12.7	11.9	11.6	10.7	10.3	10.0
	-1.28	.10	12.2	12.2	12.1	12.1	12.0	10.9	10.7	10.5	9.3	9.0	8.6	8.2	12.9	12.9	12.9	12.8	12.7	12.5	12.0	10.9	10.6	9.4	9.0	8.6
	-1.64	.05	11.6	11.6	11.5	11.4	10.1	9.9	9.7	8.2	7.9	7.6	7.2	12.4	12.4	12.4	12.3	12.2	12.1	11.5	10.1	9.8	8.3	8.0	7.6	

Note. LE = Level of education; Cum. prob. = cumulative probability.

Table A4. Normative VLT data for the Delayed recall score stratified by Level of education, Sex, and, Age (25, 30, ..., 80 years). The raw test score leading to a particular Z-value is given for Z-values indicating the percentiles 5, 10, 20, 50, 80, 90, and 95.

			MALE															FEMALE																																																																																																																																																																								
			Age in years															Age in years																																																																																																																																																																								
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80																																																																																																																																																																
LE low			Z	Cum. prob.	1.64	.95	142	141	139	136	133	130	126	122	117	112	106	100	150	150	150	148	145	142	138	134	129	124	118	112	1.28	.90	133	132	130	127	124	121	117	113	108	103	97	91	145	144	142	139	136	133	129	125	120	115	109	103	0.84	.80	122	121	119	116	113	110	106	102	97	92	86	80	134	133	131	128	125	122	118	114	109	104	98	92	0	.50	101	100	98	95	92	89	85	81	76	71	65	59	113	112	110	107	104	101	97	93	88	83	77	71	-0.84	.20	80	79	77	74	71	68	64	60	55	50	44	38	92	91	89	86	83	80	76	72	67	62	56	50	-1.28	.10	70	68	66	63	60	57	53	49	44	39	33	27	81	80	78	75	72	69	65	61	56	51	45	39	-1.64	.05	61	59	57	54	51	48	44	40	35	30	24	18	72	71	69	66	63	60	56	52	47	42	36	30
			Z	Cum. prob.	1.64	.95	150	149	147	145	142	138	134	130	125	120	114	108	25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80																																																																																																																																																
LE average			1.28	.90	142	140	138	136	133	129	125	121	116	111	105	99	93	88	143	141	139	137	134	130	126	122	117	112	106	100	150	150	150	148	145	141	137	133	128	123	117	112	106	100	0.84	.80	131	129	127	125	122	118	114	110	105	100	94	88	143	141	139	137	134	130	126	122	117	112	106	100	0	.50	110	108	106	104	101	97	93	89	84	79	73	67	122	120	118	116	113	109	105	101	96	91	85	79	-0.84	.20	89	87	85	83	80	76	72	68	63	58	52	46	101	99	97	95	92	88	84	80	75	70	64	58	-1.28	.10	78	76	74	72	69	65	61	57	52	47	41	35	90	88	86	84	81	77	73	69	64	59	53	47	-1.64	.05	69	67	65	63	60	56	52	48	43	38	32	26	81	79	77	75	72	68	64	60	55	50	44	38												
			Z	Cum. prob.	1.64	.95	150	150	150	149	146	142	139	134	129	124	119	112	25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80																																																																																																																																																
LE high			1.28	.90	146	144	142	140	137	133	130	125	120	115	110	103	97	92	150	150	150	150	150	149	145	142	137	132	127	122	116	111	104	99	92	147	145	143	141	138	134	131	126	121	116	111	104	0.84	.80	135	133	131	129	126	122	119	114	110	104	99	92	147	145	143	141	138	134	131	126	121	116	111	104	0	.50	114	113	110	108	105	102	98	93	89	83	78	72	126	124	122	120	117	113	110	105	101	95	90	83	-0.84	.20	93	92	89	87	84	81	77	72	68	62	57	51	105	104	101	99	96	93	89	84	80	74	69	63	58	-1.28	.10	82	81	78	76	73	70	66	61	57	51	46	40	94	93	90	88	85	82	78	73	69	63	58	52	-1.64	.05	73	72	70	67	64	61	57	52	48	42	37	31	85	84	81	79	76	73	69	64	60	54	49	43								

Note. LE = Level of education; Cum. prob. = cumulative probability.

Chapter 4

Normative data for the animal, profession and Letter *m* naming Verbal Fluency Tests for Dutch speaking participants and the effects of age, education, and sex

Wim Van der Elst, Martin P. J. Van Boxtel, Gerard J. P. Van Breukelen & Jelle Jolles

Abstract

Previous research has indicated that performance on Verbal Fluency Tests (VFTs) is influenced by language and/or culture. Consequently, normative VFT data for English-speaking people cannot be used for people for whom English is not their first language. The aim of the present study was to provide normative data for the Animal naming, Profession naming, and Letter M naming (four-letter words beginning with the letter M) VFTs for Dutch-speaking populations, based on a large sample ($N=1856$) of healthy men and women aged 24–81 years of different educational levels. The results showed that age affected the performance of all VFTs profoundly, but the age effect was not uniform: in the Profession and Letter M naming VFTs, performance was stable in young adulthood but declined strongly after age 50. In contrast, in the Animal naming VFT, performance appeared to decline linearly, starting early in life. Furthermore, males had higher scores than females on the Profession-naming VFT, and higher educated participants outperformed their lower educated counterparts on all three VFTs. Regression-based normative data were prepared for the three VFTs, and the advantages of using a regression-based normative approach instead of a traditional normative approach are discussed.

Published in the *Journal of the International Neuropsychological Society*. 2006; 12, 80-89

Introduction

In Verbal Fluency Tests (VFTs), testees are required to generate, typically within 60 seconds, as many words as possible that belong to a certain category (Lezak et al., 2004). Two types of VFTs are usually distinguished, namely phonemic VFTs (in which the participants are asked to generate words beginning with a given letter) and semantic VFTs (in which an individual is asked to recite examples of a given category, e.g. animals). From a psychometric viewpoint, VFTs have been used to investigate a variety of cognitive processes, e.g. word knowledge, access to semantic memory, long-term verbal memory, attention (Crowe, 1998; Ruff et al., 1997), speed of information processing, vocabulary size, working memory, inhibition of irrelevant words (Sergeant et al., 2002), and, executive functioning (Henry & Crawford, 2004; Mitrushina et al., 1999). A diminished VFT performance has been associated with a variety of pathological conditions, e.g. Alzheimer's disease (Pachana et al., 1996), schizophrenia (Crawford et al., 1993), Parkinson's disease (Flowers et al., 1995), depression (Videbech et al., 2003), traumatic brain injury (Raskin & Rearick, 1996), and HIV-associated dementia (Woods et al., 2004). In fact, VFT performance has been used in the differential diagnosis of Alzheimer disease and ischemic vascular dementia (Tierney et al., 2001).

To be used in a meaningful way, it is important that appropriate normative data are available for these tests. Different studies have provided normative data for VFTs for English-speaking people (for an overview, the reader is referred to Mitrushina et al., 1999), but these norms cannot be used for people who speak a different language because linguistic factors affect VFT performance. With respect to semantic VFTs, Kempler and colleagues (1998) found that Spanish speakers generated the smallest number of animal names in comparison to Chinese and English speakers, and Vietnamese speakers generated the most animal names. The researchers related these differences to differences in the length of words for animal names in these languages, with animal names being longest in Spanish and shortest in Vietnamese. Studies with phonemic VFTs on the other hand suggested that the difference in the number of generated words was related to the frequency of words beginning with the given letter in a language: Spanish speakers generate fewer words beginning with the letters F, A, and S in comparison to English speakers because these letters are less common in Spanish than in English (Loewenstein et al., 1994; Loewenstein & Rubert, 1992; Lopez & Taussig, 1991). In addition to these linguistic factors, culture-specific factors, such as the degree of familiarity of testing (Ardila, 1995), may also affect semantic and phonemic VFT performance.

The aim of the present study was to provide normative data for the Animal naming, Profession naming, and Letter M (four-letter words beginning with the letter M) naming VFTs for Dutch-speaking participants, using a large sample ($N=1,856$) of cognitively intact participants aged 24-81 years. In normative studies, it is important to identify which variables influence test-performance

(Lezak et al., 2004). Previous research suggested that phonemic fluency tends to be relatively resistant to the effects of aging (Harrison et al., 2000; Ivnik et al., 1996; Mathuranath et al., 2003; Tomer & Levin, 1993; but see Auriacombe et al., 2001), whilst semantic fluency is suggested to decline with age (Auriacombe et al., 2001; Benito-Cuadrado et al., 2002; Crossley et al., 1997; Harrison et al., 2000; Kosmidis et al., 2004; Lee et al., 2004; Mathuranath et al., 2003; Tomer & Levin, 1993). Results regarding the influence of education on VFT performance are less clear-cut, with some studies reporting little impact of education on either semantic or phonemic VFTs (Harrison et al., 2000) and others reporting effects of education on both types of VFTs (Kosmidis et al., 2004; Mathuranath et al., 2003; Ratcliff et al., 1998). Sex differences in VFT performance are generally not observed (Benito-Cuadrado et al., 2002; Harrison et al., 2000; Lee et al., 2004; Mathuranath et al., 2003), but some studies reported a modest effect of sex on the performance of semantic VFTs (Auriacombe et al., 2001; Kosmidis et al., 2004) or semantic and phonemic VFTs (Tombaugh et al., 1999). Ruff and colleagues (1996) suggested that sex might moderate the effect of education.

Thus, in the present study, we first established to what extent age, sex, and/or education influences performance on the different VFTs in our Dutch-speaking sample, in order to correct normative data for these variables.

Method

Participants

The data used for this study were derived from the Maastricht Aging Study, a prospective study of the determinants of usual cognitive aging (Jolles et al., 1995). Participants were recruited from the Registration Network Family Practices (*RegistratieNet Huisartspraktijken*, RNH), which contains the demographic and health characteristics of 80,000 people who live in the province of Limburg in the Netherlands. A total of 10,396 individuals aged between 24 and 81 years were randomly drawn from the register and were screened for medical pathology known to interfere with cognition (cerebrovascular pathology, all tumors of the nervous system, multiple sclerosis, all types of epilepsy, parkinsonism, dementia, organic psychosis, schizophrenia, affective psychosis, and, mental retardation). These people were told about the study by their general practitioner (which was expected to improve compliance) and were asked to return a prepaid postcard indicating their willingness (or not) to participate in the study. A total of 4,490 people agreed to participate in the study (3,531 individuals refused to participate and 2,375 individuals did not return the postcard). These people were screened in a semi-structured interview to check for additional exclusion criteria that were not coded in the RNH (i.e. a history of transient ischemic attacks, brain surgery, hemodialysis for renal failure, electrocon-

vulsive therapy, and, chronic psychotropic drug use), which led to the exclusion of a further 301 participants. Of the remaining 4,189 participants, 1,856 were randomly selected from 12 age categories (25 ± 1 years, 30 ± 1 years, ..., 80 ± 1 years) for participation in the study.

Not all data for the 1,856 participants were analyzed: a score below 24 on the Mini-Mental State Examination (Folstein et al., 1975), the occurrence of technical problems during test assessment, and the occurrence of more than five errors or repetitions in at least one VFT resulted in exclusion. In this way, unreliable data or data from participants who could not be regarded as ‘cognitively intact’ were not used. After these corrections – which excluded an additional 31 participants – the data of 1,825 participants were available for the analysis. Basic descriptive data for the sample are provided in Table 1. Level of education (LE) was assessed by classifying formal schooling according to a system often used in the Netherlands (De Bie, 1987) and which is comparable with the International Standard Classification of Education (UNESCO, Paris, 1976). The participants were grouped as follows: those with at most primary education (LE low), those with junior vocational training (LE average), and those with senior vocational or academic training (LE high). These three levels of education corresponded with an average ($\pm$ SD) of 8.59 ± 1.92 , 11.39 ± 2.50 , and 15.23 ± 3.31 years of full-time education, respectively. Using LE as a measurement of education instead of years of education has the advantage that qualitative considerations are taken into account in addition to quantitative considerations (Houx et al., 1993). All participants were native Dutch speakers and all were of Caucasian background.

Table 1. Descriptive characteristics of the sample (N=1,825).

Age		Age		Level of Education Frequency			Male:Female
Group	N	Mean	SD	Low	Average	High	Ratio
25 $\pm$ 1 year	160	25.41	.90	21	79	60	80:80
30 $\pm$ 1 year	154	30.61	.88	22	79	53	79:75
35 $\pm$ 1 year	158	35.54	.91	32	77	49	79:79
40 $\pm$ 1 year	159	40.50	.87	38	75	46	78:81
45 $\pm$ 1 year	164	45.54	.92	51	73	40	81:83
50 $\pm$ 1 year	163	50.31	.90	52	72	39	80:83
55 $\pm$ 1 year	161	55.44	.90	83	51	27	81:80
60 $\pm$ 1 year	157	60.63	.84	74	64	19	81:76
65 $\pm$ 1 year	157	65.46	.86	84	53	20	77:80
70 $\pm$ 1 year	162	70.31	.83	86	56	20	85:77
75 $\pm$ 1 year	163	75.18	.81	83	56	24	81:82
80 $\pm$ 1 year	67	79.78	.83	44	9	14	33:34
Total	1825	51.62	16.42	670	744	411	915:910

Procedure and Outcome measures

All participants were individually tested at the neuropsychological test laboratory of the University Hospital Maastricht. In all cases, the participants were asked to name as many words as possible belonging to a restricted category in 60 seconds. Three VFTs were administered in a fixed order, one after the other, namely the Animal naming, Profession naming, and Letter M naming (four-letter words beginning with the letter M) VFTs. The total number of correct, nonrepeated words was scored for each VFT.

The Animal, Profession, and letter M naming VFTs were used in the present study because these VFTs are frequently administered in the Netherlands (see e.g. Hurks et al., 2004; Luteijn & Van der Ploeg, 1983; Schmitt et al., 2000). In the letter M naming VFT, we added an extra search criterion (four-letter words; see also Thurstone, 1938) to the one that is usually evoked in standard phonemic VFTs (words beginning with a certain letter). This was done because the focus of the Maastricht Aging Study is primarily on age-related changes in cognition (Jolles et al., 1995), and previous research with standard phonemic VFTs showed that performance was relatively resistant to the effects of aging (see above). As was observed with many neuropsychological tests, age differences in performance increase as the complexity of the task increases (Salthouse, 1992, 1996). Thus, in order to increase the probability of observing age-related differences in letter M naming performance, we increased the task complexity by adding the four-letter search criterion.

The tests were administered by five test assistants who had been intensively trained in test administration by the neuropsychologists and physician of the project staff. Members of the project staff visited the test assistants at least once a week to ensure uniform test administration and data collection.

Data Analysis

Multiple linear regression models were fitted for the VFT scores using a step-down hierarchical procedure with as predictors Age, Age², Sex, LE (dummy coded with two dummies and LE average as the reference category), including all possible two-way interactions. Age was centered (Age = calendar age - 50) before computing the quadratic terms and interactions to avoid multicollinearity (Marquardt, 1980). The dummies LE low and LE high were always either both included, or both excluded from, the model, since they belong together and represent the effect of the categorical predictor level of education. Similarly, their interactions with another predictor were always either both included or excluded from the model. Non-significant predictors ($p > .005$) were excluded from the model, but no predictor was removed from the model as long as it was also included in a higher-order term in the model. In particular, Age was never removed if Age² or any interaction involving Age or Age² was still in the model. The reason for this is that the p -value of any predictor is arbitrary

(depending on the coding used for the predictors) if that predictor is part of any higher-order predictor in the model (Aiken & West, 1991). Note that a lower *alpha* level ($p = .005$) was chosen in these analyses to avoid Type I errors due to multiple testing. The assumptions of regression analysis were tested for each model: homoscedasticity (by visual inspection of the scatter plots of the residuals on the predicted values), normal distribution of the residuals (by visual inspection of the normal probability plots), multicollinearity (by calculating the Variance Inflation Factors, which should not exceed 10; Belsley et al., 1980), and influential cases (by calculating Cook's distances).

The normative data can be obtained by calculating the residuals for each VFT score ($e_i = \text{observed score} - \text{predicted score}$): after standardization of the residuals ($Z_i = e_i / \text{SD}[\text{residual}]$), the performance of the participant can be evaluated via a Z distribution table with cumulative probabilities. More details regarding the regression-based normative method can be found elsewhere (Van Breukelen & Vlaeyen, in press). All analyses were performed using the SPSS 10.0 for Macintosh software package.

Results

Table 2 presents the regression models for the Animal, Profession, and Letter M naming VFTs. The assumptions of regression analysis were fulfilled for each model (no heteroscedasticity, normally distributed residuals, maximum VIF = 1.301, and maximum Cook's distance = .012). Age negatively influenced performance on all three VFTs, with an additional significant quadratic Age-effect on the Profession and Letter M naming VFTs. Sex differences were only found in the Profession naming VFT, with males performing better than females. Higher educated participants outperformed their lower educated counterparts on all VFTs. The negative impact of a low education as compared with an average education was larger than the positive effect of a high education for all VFT scores (see standardized B values, Table 2). None of the interaction terms were significant.

The regression models, combined with the standard deviations of the residuals (see Table 2, $\text{SD}[\text{residual}]$), provide normative data. First, the participant's predicted VFT scores are calculated. Then the residuals of these scores are computed ($e_i = \text{observed score} - \text{predicted score}$) and standardized ($Z_i = e_i / \text{SD}[\text{residual}]$), so that the individual's performance can be evaluated using a standard Z distribution table with corresponding cumulative probabilities.

For example, in the Profession naming VFT, a 60-year-old woman with a high LE generated 15 profession names in 1 minute. Her predicted Profession naming score would be $19.498 (= 18.579 - (.054*10) - (.002*100) + (.753*0) - (2.742*0) + (1.659*1))$, with a residual of $-4.498 (= 15 - 19.498)$. The standardized residual equals $-.934 (= -4.498/4.818)$, which corresponds to a p value of .175.

Table 2. Final multiple linear regression models of the VFT measures resulting from a step-down hierarchical procedure; the full model included Age, Age², Sex, LE low, LE high, and all two-way interactions as predictors.

VFT score	Variable	B	Std. Error B	T	Standardized B	R ²	SD(residual)
Animal naming	(constant)	24.777	.213	116.210*			
	Age	-.097	.009	-11.065*	-.249		
	LE low	-2.790	.321	-8.688*	-.210		
	LE high	1.586	.358	4.4321*	.104	.177	5.797
Profession naming	(constant)	18.579	.243	76.317*			
	Age	-.054	.007	-7.259*	-.166		
	Age ²	-.002	< .001	-3.969*	-.086		
	Sex	.753	.228	-3.301*	.070		
	LE low	-2.742	.267	-10.255*	-.247		
	LE high	1.659	.299	5.547*	.129	.189	4.818
Letter M naming	(constant)	10.588	.164	64.481*			
	Age	-.035	.006	-6.084*	-.138		
	Age ²	-.001	< .001	-3.320*	-.071		
	LE low	-2.331	.204	-11.449*	-.274		
	LE high	1.587	.227	6.993*	.162	.198	3.675

Note. LE = Level of education. Encoding of the predictors: Age = calendar age - 50; Age² = (calendar age - 50)²; Sex: Male = 1, Female = 0; LE Low: Low Education = 1, Average or High Education = 0; LE High: High Education = 1, Low or Average Education = 0.

* $p < .001$

Tables A1, A2, and A3 (in Appendix) provide normative data for the VFTs based on the regression models, stratified by their statistically relevant predictors (see Table 2). If an individual is not exactly 25, 30, ..., 80 years old, then the person's age should be rounded up to the closest age given. If the performance of such an individual is borderline-normal according to Tables A1, A2, and A3 (e.g. Z value ≈ -1.28), the regression models presented in Table 2 can be used to determine the exact Z values.

Discussion

In the present study, we first investigated the influence of Age, Sex, and LE on VFT performance. Although previous research has shown that semantic fluency declines with age whereas phonemic fluency remains quite stable, our study suggests that age affects performance on both types of VFTs profoundly. The age-related differences in phonemic VFT performance found in the present study may be because we added a second searching criterion – namely four-

letter words – to the one that is usually employed in phonemic VFTs (words beginning with a certain letter). Indeed, many empirical studies comparing the performance of younger versus older adults on cognitive tasks have shown that increasing task complexity leads to greater age differences in performance, a phenomenon that is known as the ‘complexity effect’ (Salthouse, 1992, 1996). As mentioned by Salthouse (1992), understanding the complexity effect is important because it may contribute to the understanding of the causes of adult age differences in cognition. Thus, phonemic VFTs with a restriction in the length of the words to be generated may be a valuable addition to the assortment of neuropsychological tests that are commonly used in aging research. An alternative explanation for the finding of age effects on the letter M naming VFT whilst age effects are usually not observed in phonemic VFTs may be the higher statistical power of the present study (large sample) in comparison to most other VFT studies. More research is needed to determine which explanation is correct, for example by using a counterbalanced design in which phonemic VFTs with and without the four-letter search criterion are administered. However, the age effects found in the present study were substantial (i.e. the predicted letter M naming score for an 80-year-old person is about 25.8%, 20.3%, and 17.7% lower than the predicted performance of a 25-year-old person for low, average, and high educated people, respectively; see Appendix Table A3), which suggest that these age effects should also be detected in smaller samples with sufficient age variability and are not primarily power-related.

In contrast to most of the other VFT studies, we used multiple regression models with both linear and quadratic age effects as predictors of VFT performance. An advantage of using such a method is that linear age effects can be distinguished from quadratic age effects: only a linear age effect was found in the Animal naming VFT, whereas significant quadratic age effects in addition to the linear age effect were observed in the Profession naming and Letter M naming VFTs. As an example, Figure 1 shows the predicted VFT scores at ages 30, 35, ..., 80 relative to the predicted VFT scores at age 25 for average educated females³. As shown, the performance on the Profession and Letter M naming VFTs appeared to remain quite stable until the age of about 50, and then declined, whereas performance on the Animal naming VFT declined in a linear way. The differential effect of age on Animal and Profession naming verbal fluency was unexpected, because both tasks are semantic VFTs, and it is generally thought that performance on different semantic VFTs (e.g. Animal and Profession naming) relies on the same brain areas and/or cognitive processes,

³ Figure 1 shows the relative predicted VFT scores (based on Table 2) rather than the raw predicted scores for reasons of clarity and comparability, because the raw predicted scores (number of generated words) differ per VFT (more Animal names are generated than Profession names, and more Profession names than Letter M words). Figure 1 presents these values for average educated females only, but the shape of these curves is very similar for people with different demographical characteristics.

e.g. on temporal lobe mediated semantic knowledge (Gleissner & Elger, 2001). Thus, the Animal and Profession naming VFTs would be expected to measure the same construct(s) and consequently both tasks should be influenced in a similar way by an important independent variable such as age. It is unclear what caused this difference in the effect of age on Animal and Profession naming performance, and this issue merits further research (e.g. using a counterbalanced order of presentation of the VFTs to rule out possible order-effects that may be responsible for this difference).

Figure 1. Predicted VFT scores at ages 30, 35, ..., and 80 years relative to the predicted VFT scores at age 25 years, for average educated females.

In agreement with most other studies, we found that the effect of sex on VFT performance was relatively small: males generated on average .753 Profession names more than females did, but there were no sex differences in the Animal or Letter M Naming VFTs. Sex did not appear to modulate the effect of education, as was suggested by Ruff and colleagues (1996). Performance on both semantic and phonemic VFTs was profoundly affected by educational attainment, with there being a greater difference in VFT performance between low and average educated people than between average and high educated individuals. For the Letter M naming and Profession naming VFTs, a low LE was the most important predictor of test performance (compare the standardized B's, Table 2). For the Animal naming VFT, age was the most important predictor, but LE also influenced performance profoundly, i.e. low-educated participants generated on average 2.790 fewer words than average-educated participants and

4.376 fewer words than high-educated participants (see Table 2), an effect that corresponds with the effect of about 28.8 ($= 2.790/.097$) and 45.1 ($= 4.376/.097$) years of aging, respectively. As indicated by Mitrushina and colleagues (1999), nearly every normative VFT study with English-speaking people is confined to participants with a high level of educational attainment. Our finding that especially a low LE adversely influenced VFT performance is of importance in this respect because it suggests that the available normative data for English-speaking participants may seriously bias the evaluation of the VFT performance of lower educated people. This is especially problematic because a significant proportion of the people who are tested neuropsychologically are older persons, and many older people are relatively poorly educated (before the 1960s most people did not go on to higher education, often because of socio-economical rather than intellectual reasons; Jolles et al., 1995). This decrease in educational attainment as a function of increasing age was also seen in our sample (see Table 1). An advantage of using a regression-based normative approach is that such unbalanced data can be analyzed without biasing the estimation or testing of the regression weights, and with only a minor loss of statistical power (because the standard error of a regression weight is proportional to the $\sqrt{\text{VIF}}$ of the predictor at hand; see e.g. Kleinbaum, Kupper, Muller & Nizam, 1998). Indeed, regression-based norms provide accurate estimates of the population statistics because they are based on equations that are derived using the data for all demographic groups (Van Breukelen & Vlaeyen, in press; Zachary & Gorsuch, 1985). This is not the case with a traditional normative approach, which provides means and SDs of the test scores of the normative sample, so that a tested person's raw score can be evaluated by converting it into a Z-score. A problem that is intrinsically related to the traditional method is that – even when the overall size of the normative sample (N) is reasonably large – the sample size of the actual reference group (n) is small after the sample is broken down by the relevant demographic variables (Crawford & Howell, 1998). For example, splitting a sample by age group (12 levels), sex (2 levels), and LE (3 levels) reduces the sample size per subgroup to $1/72^{\text{th}}$ of the total sample size. As a result, the population statistics (mean, SD) can be much less accurately estimated within subgroups than for the total sample, which yields unreliable normative data (Van Breukelen & Vlaeyen, in press).

Age, Sex, and LE are important factors that are related to VFT performance (Mitrushina et al., 1999), but previous studies showed that other factors such as verbal intelligence also affect VFT performance (Bolla et al., 1998). In fact, Bolla et al. (1998) found that verbal intelligence has a stronger association with VFT performance than education has. Nevertheless, we used LE as a proxy of 'general ability' rather than verbal IQ for a number of reasons. First, it is important that an independent variable can be measured in a straightforward manner in normative studies. Using verbal intelligence as a predictor of VFT performance would imply that a person would have to complete a verbal intelligence

test before the normative VFT data could be used, which is a time-consuming and expensive process (Van der Elst et al., 2005). Secondly, it is important that independent variables used in a regression model are measured reliably. Although an independent variable such as LE is somehow imperfectly measured because it is usually based on self-report (which was also the case in the present study), the measurement error associated with a measure such as IQ is substantially larger. A large measurement error makes the independent variable an imperfect statistical control and can invalidate a regression model (Fox, 1997), and thus the derived normative data. Thirdly, in normative studies, education is the preferred operationalization of ‘general ability’. Using a measure such as LE instead of verbal intelligence has the advantage of a better comparability with previous normative research. Fourthly, when IQ is used as a stratification variable, certain specific problems, such as the Flynn effect (the gain of IQ over time; Flynn, 1999), are encountered. As a result, researchers or clinicians using the VFT normative data would have to use the same IQ norms as the ones used in the specific normative study, even though more up-to-date IQ norms are available.

It should be noted that we used a cross-sectional design in the present study, which has certain limitations. For example, with such a study design it cannot be determined to what extent the effects of age on VFT performance are confounded by cohort effects. However, this was not a major problem in our study because we were primarily interested in normative trends in VFT performance, and a cross-sectional study design is appropriate for such purposes (Hayslip & Panek, 1989). An advantage of the present study is the very large number of healthy, cognitively intact participants from whom normative data were derived using a regression-based approach. The normative data can be used to evaluate the Animal, Profession, and Letter M naming VFT performance of Dutch-speaking participants aged 24–81 years old, with appropriate correction for demographic influences.

Appendix

Table A1. Normative data for the Animal naming VFT, stratified by Age (25, 30, ..., 80 years) and Level of Education. The raw score leading to a particular Z-value is given for Z-values indicating the percentiles 5, 10, 20, 50, 80, 90 and 95.

	Z value	Cum. prob.	MALE & FEMALE											
			Age in years											
			25	30	35	40	45	50	55	60	65	70	75	80
LE low	1.64	.95	33.9	33.4	32.9	32.5	32.0	31.5	31.0	30.5	30.0	29.6	29.1	28.6
	1.28	.90	31.8	31.3	30.9	30.4	29.9	29.4	28.9	28.4	28.0	27.5	27.0	26.5
	0.84	.80	29.3	28.8	28.3	27.8	27.3	26.9	26.4	25.9	25.4	24.9	24.4	23.9
	0	.50	24.4	23.9	23.4	23.0	22.5	22.0	21.5	21.0	20.5	20.0	19.6	19.1
	-0.84	.20	19.5	19.1	18.6	18.1	17.6	17.1	16.6	16.1	15.7	15.2	14.7	14.2
	-1.28	.10	17.0	16.5	16.0	15.5	15.1	14.6	14.1	13.6	13.1	12.6	12.1	11.7
	-1.64	.05	14.9	14.4	13.9	13.4	13.0	12.5	12.0	11.5	11.0	10.5	10.1	9.6
LE average	Z value	Cum. prob.	25	30	35	40	45	50	55	60	65	70	75	80
	1.64	.95	36.7	36.2	35.7	35.3	34.8	34.3	33.8	33.3	32.8	32.3	31.9	31.4
	1.28	.90	34.6	34.1	33.7	33.2	32.7	32.2	31.7	31.2	30.7	30.3	29.8	29.3
	0.84	.80	32.1	31.6	31.1	30.6	30.1	29.6	29.2	28.7	28.2	27.7	27.2	26.7
	0	.50	27.2	26.7	26.2	25.7	25.3	24.8	24.3	23.8	23.3	22.8	22.4	21.9
	-0.84	.20	22.3	21.8	21.4	20.9	20.4	19.9	19.4	18.9	18.5	18.0	17.5	17.0
	-1.28	.10	19.8	19.3	18.8	18.3	17.8	17.4	16.9	16.4	15.9	15.4	14.9	14.4
LE high	Z value	Cum. prob.	25	30	35	40	45	50	55	60	65	70	75	80
	1.64	.95	38.3	37.8	37.3	36.8	36.4	35.9	35.4	34.9	34.4	33.9	33.4	33.0
	1.28	.90	36.2	35.7	35.2	34.8	34.3	33.8	33.3	32.8	32.3	31.8	31.4	30.9
	0.84	.80	33.7	33.2	32.7	32.2	31.7	31.2	30.7	30.3	29.8	29.3	28.8	28.3
	0	.50	28.8	28.3	27.8	27.3	26.8	26.4	25.9	25.4	24.9	24.4	23.9	23.5
	-0.84	.20	23.9	23.4	22.9	22.5	22.0	21.5	21.0	20.5	20.0	19.6	19.1	18.6
	-1.28	.10	21.4	20.9	20.4	19.9	19.4	18.9	18.5	18.0	17.5	17.0	16.5	16.0
	Z value	Cum. prob.	25	30	35	40	45	50	55	60	65	70	75	80
	1.64	.95	19.3	18.8	18.3	17.8	17.3	16.9	16.4	15.9	15.4	14.9	14.4	13.9
	1.28	.90												
	0.84	.80												
	0	.50												
	-0.84	.20												
	-1.28	.10												

Note. LE = Level of education; Cum. prob. = cumulative probability.

Table A2. Normative data for the Profession naming VFT, stratified by Age (25, 30, ..., 80 years), Sex, and Level of Education. The raw score leading to a particular Z-value is given for Z-values indicating the percentiles 5, 10, 20, 50, 80, 90 and 95.

	Z	Cum. value	MALE															FEMALE														
			Age in years															Age in years														
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80						
LE low	1.64	.95	24.6	24.8	24.9	24.8	24.7	24.5	24.2	23.8	23.2	22.6	21.9	21.1	21.1	23.8	24.0	24.1	24.1	24.0	23.7	23.4	23.0	22.5	21.9	21.1	20.3					
	1.28	.90	22.9	23.0	23.1	23.1	23.0	22.8	22.4	22.0	21.5	20.9	20.2	19.3	19.3	22.1	22.3	22.4	22.3	22.2	22.0	21.7	21.3	20.7	20.1	19.4	18.6					
	0.84	.80	20.7	20.9	21.0	21.0	20.9	20.6	20.3	19.9	19.4	18.8	18.0	17.2	17.2	20.0	20.2	20.2	20.2	20.1	19.9	19.6	19.1	18.6	18.0	17.3	16.5					
	0	.50	16.7	16.9	17.0	16.9	16.8	16.6	16.3	15.9	15.3	14.7	14.0	13.2	13.2	15.9	16.1	16.2	16.2	16.1	15.8	15.5	15.1	14.6	14.0	13.2	12.4					
	-0.84	.20	12.6	12.8	12.9	12.9	12.8	12.5	12.2	11.8	11.3	10.7	9.9	9.1	9.1	11.9	12.1	12.1	12.1	12.0	11.8	11.5	11.0	10.5	9.9	9.2	8.4					
	-1.28	.10	10.5	10.7	10.8	10.8	10.6	10.4	10.1	9.7	9.2	8.5	7.8	7.0	7.0	9.8	9.9	10.0	10.0	9.9	9.7	9.3	8.9	8.4	7.8	7.1	6.2	5.4				
	-1.64	.05	8.8	9.0	9.0	9.0	8.9	8.7	8.4	7.9	7.4	6.8	6.1	5.3	5.3	8.0	8.2	8.3	8.3	8.2	7.9	7.6	7.2	6.7	6.1	5.3	4.5					
LE average	Z	Cum. value	25	30	35	40	45	50	55	60	65	70	75	80	80	25	30	35	40	45	50	55	60	65	70	75	80					
	1.64	.95	27.3	27.5	27.6	27.6	27.5	27.2	26.9	26.5	26.0	25.4	24.6	23.8	23.8	26.6	26.8	26.8	26.8	26.7	26.5	26.2	25.7	25.2	24.6	23.9	23.1					
	1.28	.90	25.6	25.8	25.9	25.8	25.7	25.5	25.2	24.8	24.2	23.6	22.9	22.1	22.1	24.8	25.0	25.1	25.1	25.0	24.7	24.4	24.0	23.5	22.9	22.1	21.3					
	0.84	.80	23.5	23.7	23.7	23.7	23.6	23.4	23.1	22.6	22.1	21.5	20.8	20.0	20.0	22.7	22.9	23.0	23.0	22.8	22.6	22.3	21.9	21.4	20.7	20.0	19.2					
	0	.50	19.4	19.6	19.7	19.7	19.6	19.3	19.0	18.6	18.1	17.5	16.7	15.9	15.9	18.7	18.9	18.9	18.9	18.8	18.6	18.3	17.8	17.3	16.7	16.0	15.2					
	-0.84	.20	15.4	15.6	15.6	15.6	15.5	15.3	15.0	14.5	14.0	13.4	12.7	11.9	11.9	14.6	14.8	14.9	14.9	14.8	14.5	14.2	13.8	13.3	12.7	11.9	11.1					
	-1.28	.10	13.3	13.4	13.5	13.5	13.4	13.2	12.8	12.4	11.9	11.3	10.6	9.7	9.7	12.5	12.7	12.8	12.8	12.6	12.4	12.1	11.7	11.2	10.5	9.8	9.0					
-1.64	.05	11.5	11.7	11.8	11.8	11.7	11.4	11.1	10.7	10.2	9.6	8.8	8.0	8.0	10.8	11.0	11.0	11.0	10.9	10.7	10.4	9.9	9.4	8.8	8.1	7.3						
LE high	Z	Cum. value	25	30	35	40	45	50	55	60	65	70	75	80	80	25	30	35	40	45	50	55	60	65	70	75	80					
	1.64	.95	29.0	29.2	29.3	29.2	29.1	28.9	28.6	28.2	27.6	27.0	26.3	25.5	25.5	28.2	28.4	28.5	28.5	28.4	28.1	27.8	27.4	26.9	26.3	25.5	24.7					
	1.28	.90	27.3	27.4	27.5	27.5	27.4	27.2	26.8	26.4	25.9	25.3	24.6	23.7	23.7	26.5	26.7	26.8	26.7	26.6	26.4	26.1	25.7	25.1	24.5	23.8	23.0					
	0.84	.80	25.1	25.3	25.4	25.4	25.3	25.0	24.7	24.3	23.8	23.2	22.4	21.6	21.6	24.4	24.6	24.6	24.6	24.5	24.3	24.0	23.5	23.0	22.4	21.7	20.9					
	0	.50	21.1	21.3	21.4	21.3	21.2	21.0	20.7	20.3	19.7	19.1	18.4	17.6	17.6	20.3	20.5	20.6	20.6	20.5	20.2	19.9	19.5	19.0	18.4	17.6	16.8					
	-0.84	.20	17.0	17.2	17.3	17.3	17.2	16.9	16.6	16.2	15.7	15.1	14.3	13.5	13.5	16.3	16.5	16.6	16.6	16.5	16.4	16.2	15.9	15.5	14.9	14.3	13.6	12.8				
	-1.28	.10	14.9	15.1	15.2	15.2	15.0	14.8	14.5	14.1	13.6	12.9	12.2	11.4	11.4	14.2	14.4	14.4	14.4	14.3	14.1	13.8	13.3	12.8	12.2	11.5	10.7					
-1.64	.05	13.2	13.4	13.4	13.4	13.3	13.1	12.8	12.3	11.8	11.2	10.5	9.7	9.7	12.4	12.6	12.7	12.7	12.6	12.3	12.0	11.6	11.1	10.5	9.7	8.9						

Note. LE = Level of education; Cum. prob. = cumulative probability.

Table A3. Normative data for the Letter M naming VFT, stratified by Age (25, 30, ..., 80 years) and Level of Education. The raw score leading to a particular Z-value is given for Z-values indicating the percentiles 5, 10, 20, 50, 80, 90 and 95.

	Z value	Cum. Prob.	MALE & FEMALE											
			Age in years											
			25	30	35	40	45	50	55	60	65	70	75	80
LE low	1.64	.95	14.5	14.6	14.6	14.5	14.4	14.3	14.1	13.8	13.5	13.2	12.8	12.3
	1.28	.90	13.2	13.3	13.3	13.2	13.1	13.0	12.8	12.5	12.2	11.9	11.5	11.0
	0.84	.80	11.6	11.6	11.6	11.6	11.5	11.3	11.1	10.9	10.6	10.2	9.8	9.4
	0	.50	8.5	8.6	8.6	8.5	8.4	8.3	8.1	7.8	7.5	7.2	6.8	6.3
	-0.84	.20	5.4	5.5	5.5	5.4	5.3	5.2	5.0	4.7	4.4	4.1	3.7	3.2
	-1.28	.10	3.8	3.9	3.9	3.8	3.7	3.6	3.4	3.1	2.8	2.5	2.1	1.6
	-1.64	.05	2.5	2.5	2.5	2.5	2.4	2.2	2.0	1.8	1.5	1.1	0.7	0.3
LE average	Z value	Cum. prob.	Age in years											
			25	30	35	40	45	50	55	60	65	70	75	80
	1.64	.95	16.9	16.9	16.9	16.9	16.8	16.6	16.4	16.2	15.9	15.5	15.1	14.7
	1.28	.90	15.5	15.6	15.6	15.5	15.4	15.3	15.1	14.8	14.5	14.2	13.8	13.3
	0.84	.80	13.9	14.0	14.0	13.9	13.8	13.7	13.5	13.2	12.9	12.6	12.2	11.7
	0	.50	10.8	10.9	10.9	10.8	10.7	10.6	10.4	10.1	9.8	9.5	9.1	8.6
	-0.84	.20	7.8	7.8	7.8	7.8	7.7	7.5	7.3	7.1	6.8	6.4	6.0	5.6
LE high	-1.28	.10	6.1	6.2	6.2	6.1	6.0	5.9	5.7	5.4	5.1	4.8	4.4	3.9
	-1.64	.05	4.8	4.9	4.9	4.8	4.7	4.6	4.4	4.1	3.8	3.5	3.1	2.6
	Z value	Cum. prob.	Age in years											
			25	30	35	40	45	50	55	60	65	70	75	80
	1.64	.95	18.5	18.5	18.5	18.5	18.4	18.2	18.0	17.8	17.5	17.1	16.7	16.3
	1.28	.90	17.1	17.2	17.2	17.1	17.0	16.9	16.7	16.4	16.1	15.8	15.4	14.9
	0.84	.80	15.5	15.6	15.6	15.5	15.4	15.3	15.1	14.8	14.5	14.2	13.8	13.3
LE high	0	.50	12.4	12.5	12.5	12.4	12.3	12.2	12.0	11.7	11.4	11.1	10.7	10.2
	-0.84	.20	9.3	9.4	9.4	9.3	9.2	9.1	8.9	8.6	8.3	8.0	7.6	7.1
	-1.28	.10	7.7	7.8	7.8	7.7	7.6	7.5	7.3	7.0	6.7	6.4	6.0	5.5
	-1.64	.05	6.4	6.4	6.4	6.4	6.3	6.1	5.9	5.7	5.4	5.0	4.6	4.2

Note. LE = Level of education; Cum. prob. = cumulative probability.

Chapter 5

The Concept Shifting Test: adult normative data

Wim Van der Elst, Martin P. J. Van Boxtel, Gerard J. P. Van Breukelen & Jelle Jolles

Abstract

The Concept Shifting Test (CST) is a newly developed Trail Making Type test that measures concept shifting and executive functioning. The purpose of the present study was to determine whether CST performance was affected by age, gender, educational level, and/or handedness, and to establish the normal range of test performance. The CST was administered to a large sample ($N=1,794$) of cognitively intact adults (age range 24 to 81 years). The results showed that CST performance was superior in women, decreased with age, and increased with level of education. Handedness did not affect CST performance. Normative data that are corrected for the influences of these demographical variables were established using a regression-based normative approach.

Introduction

The well-known Trail Making Test (TMT) is an extensively used neuropsychological test that has been found to distinguish normal control participants from people with various types of brain injury (for a review, see Lezak, Howieson, & Loring, 2004). The basic rationale behind the TMT's test paradigm is that Part B (which involves alternating between numbers and letters) differs from Part A (which involves number tracing) only with respect to the extra requirement to alternate between groups of stimuli. It is reasoned that the extra time needed to complete Part B of the TMT relative to Part A reflects the 'cost' of switching between numbers and letters (Houx, 1991). An exaggeratedly slower performance on Part B relative to Part A is assumed to be indicative of impaired executive functioning (Pontius & Yudowitz, 1980). However, different authors have pointed out that some aspects of the TMT's test design could be improved (Gaudino, Geisler, & Squires, 1995; Salthouse et al., 2000). For example, the total distance between the targets in Part B of the TMT is substantially larger than in Part A (Gaudino et al., 1995), which means that the two test parts are not comparable in terms of the motor performance needed to complete the tests. Also, Part A of the TMT uses numbers whilst Part B uses numbers as well as letters as targets, which makes the subtasks not comparable with respect to the long-term memory processes involved in test performance (i.e. knowledge of the order of numbers in Part A versus knowledge of the order of numbers and knowledge of the alphabet in Part B). As a result, abnormal slowing in TMT Part B relative to Part A may be caused by poor psychomotor abilities or by poor knowledge of the alphabet rather than by a concept shifting or executive function problem. While it is impossible to develop a (neuro)psychological test in which there is a perfect one-to-one relationship between the test score and the underlying attribute to be measured, some of the problems of the TMT that complicate the interpretation of the executive function measure can be avoided. To this end, the Concept Shifting Test (CST) was developed, which has a conceptual background comparable to that of the TMT (Jolles, Houx, van Boxtel, & Ponds, 1995; Vink & Jolles, 1985).

The CST consists of four basic parts. The layout of the stimulus material is identical in each part, with 16 small circles that are grouped in a larger circle (see Figure 1). The small circles contain digits, letters, both digits and letters, or are empty. In three consecutive test trials, the participants have to cross out as quickly as possible the digits in numerical order (CST Part A), the letters in alphabetical order (CST Part B), and the digits and letters in alternating order (CST Part C). Next, there is a zero condition in which the participants are required to cross out the empty circles as fast as possible in a clockwise fashion (CST zero). The CST zero trial is done twice and serves as a correction for basic motor speed for CST Parts A, B, and C, so that the test performance in these trials can be related more unequivocally to the cognitive processes involved in the task rather than to the motor components. These cognitive processes differ

in CST Parts A, B, and C. The extra time needed to complete CST Part A and CST Part B compared with CST zero trials is thought to be mainly attributable to the factors of attention, visual recognition (of digits and letters), long-term memory (knowledge of the order of numbers and the alphabet), working memory (for keeping track of the ongoing conduct) and visual scanning (searching the target item). These factors also influence performance on CST Part C, but this subtask also requires the additional shifting of the stimulus sets, or concepts, namely letters and numbers (Houx, 1991). Thus, by correcting the time needed to complete CST Part C for the time needed to complete CST Parts A and B, a measure of concept shifting or executive functioning is obtained.

In order to make the CST widely available for clinical and research purposes, the present study's goal is to determine whether age, gender, educational level and/or handedness affects task performance in a large community sample of adults to establish the normal range of performance. A regression-based normative approach was used to establish the norms because this method has some important methodological advantages over a traditional normative approach (see Van Breukelen & Vlaeyen, 2005; Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2006a).

Method

Participants

This study was part of a larger prospective study of the determinants of normal cognitive aging, the Maastricht Aging Study (MAAS; Jolles et al., 1995). In MAAS, four panels of cognitively intact participants (1,856 individuals in total) were tested at baseline, using a battery of neuropsychological tests including the CST. The order in which the tests were administered was different for each panel of participants (see Jolles et al., 1995), but preliminary analyses showed that CST scores were not significantly different between the four panels (data not shown), which justifies merging of the data.

The data of participants who scored less than 24 on the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975) were excluded from data analysis, since these participants were considered to be at risk of cognitive disorders according to clinical guidelines. The data of participants who did not comply with the test instructions or who could not perform the CST adequately due to physical or technical problems during test assessment were also excluded. Thus the CST data of 1,794 participants were analyzed. Basic descriptive data for the normative sample are provided in Table 1. Level of education (LE) was assessed by classifying formal schooling in a system often used in the Netherlands (De Bie, 1987), which is comparable with the International Standard Classification of Education (UNESCO, Paris, 1976). Three groups were formed – those with at most primary education (LE low), those with junior vocational

training (LE average), and those with senior vocational or academic training (LE high). These three levels of education corresponded with a mean ($\pm SD$) of 8.60 (± 1.83), 11.41 (± 2.49), and 15.25 (± 3.31) years of full-time education, respectively, in this sample. Most (95.7%) of the participants were right-handed (as measured with the Edinburgh Handedness Inventory; Oldfield, 1971). All participants were Caucasian, and all were native Dutch speakers. More details regarding the sample frame, subject inclusion procedure, stratification criteria, and other aspects of the MAAS study can be found elsewhere (Jolles et al., 1995).

Figure 1. Stimulus material for the CST Part A [a], CST Part B [b], CST Part C [c] and CST zero [d] test trials. The stimulus material is presented here on a reduced scale. The true diameters of the small and larger circles are 0.59 inch [1.5 cm] and 6.29 inch [16 cm] respectively.

Table 1. Descriptive characteristics of the sample that was given the Concept Shifting Test (N=1,794).

Age Group	N	Age		Level of Education Frequency			Male:Female Ratio
		Mean	SD	Low	Average	High	
25±1 year	160	25.4	0.90	20	79	61	79:81
30±1 year	151	30.6	0.88	19	79	53	77:74
35±1 year	159	35.5	0.91	33	77	49	79:80
40±1 year	157	40.5	0.87	37	74	46	76:81
45±1 year	161	45.5	0.92	48	73	40	78:83
50±1 year	159	50.3	0.90	48	72	39	79:80
55±1 year	159	55.4	0.90	81	51	27	79:80
60±1 year	155	60.6	0.83	70	65	19	79:76
65±1 year	154	65.5	0.87	82	52	20	76:78
70±1 year	157	70.3	0.82	81	56	20	83:74
75±1 year	156	75.2	0.80	77	55	23	80:76
80±1 year	66	79.8	0.83	40	10	15	32:34
Total	1,794	51.4	16.41	636	743	412	897:897

Note. Data on level of education were missing for three participants.

Procedure and Instruments

All participants were tested individually at the neuropsychological laboratory of the Brain & Behaviour Institute (Maastricht, Netherlands). The CST test materials consist of five sheets of white paper showing 16 small circles (diameter 0.59 inch [1.5 cm]) arranged in a larger circle (diameter 6.29 inch [16 cm]). The small circles contain digits, letters, both digits and letters, or are empty (see Figure 1). In three consecutive test trials the participants were instructed to cross out as quickly as possible the fixed randomly arranged numbers 1-16 in numerical order (CST Part A), the fixed randomly arranged letters A-P in alphabetical order (CST Part B), and, the fixed randomly arranged numbers (1 to 8) alternating with letters (A to H) in the appropriate order (1, A, 2, B etc.; CST Part C). Before each test trials, a practice trial was administered to make sure that participants understood the test instructions (the practice trials consisted of cards with only 6 stimuli). Completion of the test trials was immediately followed by two zero trials, in each of which the participants had to cross out the 16 empty circles clockwise as quickly as possible. The two zero trials were administered after the CST test trials because the instruction to cross out the empty circles in a clockwise fashion could trigger a clockwise visual scanning strategy that is retained in the CST test trials.

The average time to complete the two zero trials served as an index of basic motor speed (basic motor speed = [time to complete CST zero trial 1 + time to

complete CST zero trial 2] / 2). Outcome was the time needed to complete each CST test trial corrected for basic motor speed ($CSTa$ = time to complete Part A - basic motor speed; $CSTb$ = time to complete Part B - basic motor speed; $CSTc$ = time to complete Part C - basic motor speed). The Shifting score was calculated by subtracting the average $CSTa$ and $CSTb$ scores from the $CSTc$ score ($Shifting = CSTc - [CSTa + CSTb] / 2$; Valentijn et al., 2005). In addition, the number of errors made in each CST part was recorded to take possible speed-accuracy tradeoffs into account (Errors $CSTa$, Errors $CSTb$ and Errors $CSTc$ respectively). The examiners did not point out the errors made during the test, and the participants were not required to correct errors. This avoids the most common problem in TMT testing, namely that of quickly correcting the test subject's errors so that there is no undue penalization in terms of the timed scores (Jarvis & Barth, 1987). Errors were scored when a participant did not start with the appropriate beginning item, when items were skipped, and when items were not crossed out in the appropriate order. For example, if a participant crossed out the sequence 'A, 1, B, 2, C, 4, D, 5, E, 6, F, 7, G, 8, H' in CST Part C, two errors were scored – i.e. one error for not beginning with the appropriate beginning item (the participant started with a letter instead of with a number), and one error for skipping the number 3 (2, C, 4, D, ... instead of 2, C, 3, D, ...). The Error $CSTa$, Error $CSTb$ and Error $CSTc$ scores are referred to as 'accuracy measures', and the $CSTa$, $CSTb$, $CSTc$ and Shifting scores as 'speed measures'.

Statistical Analyses

Multiple linear or binary logistic regression analysis was used to determine which variables were predictive of the CST scores. The level of α error was set to .005 for all analyses to correct for multiple testing (Bonferroni correction for the number of dependent variables, critical $p = .05 / 7 \approx .005$). All analyses were performed using the SPSS 11.5 for Windows software package. The distribution of the accuracy measures was extremely skewed and they were confined to a small part of the possible values (i.e. reversed *J*-shape distributed with 98.6%, 89.3% and 71.7% of the total sample scoring zero on the Error $CSTa$, Error $CSTb$, and Error $CSTc$ measures respectively). Analyzing such data with linear regression analysis would violate the distributional assumptions of normally distributed residuals and homoscedasticity. Therefore the Error scores were analyzed with binary logistic regression analysis after dichotomizing the raw error scores (with 'one or more Errors' = 1 and 'zero Errors' = 0). The speed measures (measured in seconds) are continuous variables with a positively skewed distribution. These variables were analyzed with multiple linear regression analysis after a square root transformation, because preliminary data analyses showed that the residuals were positively skewed before transformation of the scores (data not shown). A positive constant of 10 was added to the Shifting

score before it was square root transformed ($=\sqrt{[\text{Shifting} + 10]}$), because 3.2% of the normative sample had a negative Shifting score (see the Discussion).

The CST scores were regressed on age, age² (which enables modeling of quadratic age associations), gender, level of education, handedness, and all possible two-way interactions (the full models). Age was centred (age = calendar age - 50) before computing the quadratic terms and interactions to avoid multicollinearity (Marquardt, 1980). Gender was dummy coded with male = 1 and female = 0. Level of education (LE) was dummy coded with two dummies (LE low and LE high) and LE average as the reference category. Handedness was dummy coded with right-handed = 0 and left-handed = 1. The full models were then reduced in a stepwise way by eliminating the least significant predictor if its two-tailed *p*-value was above .005. The dummies LE low and LE high were always either both included in or both left out of the model since they belong together and represent the effect of the categorical predictor education. Similarly, their interactions with another predictor were always either both included in or both excluded from the model. In addition, a predictor was not removed from the model if it was included in a higher-order term, because the *p*-value of any predictor is arbitrary (depending on the coding used for predictors) if that predictor is part of a higher-order predictor in the model.

Regarding the multiple linear regression models, we tested the assumptions of homoscedasticity, normal distribution of the residuals, absence of multicollinearity, and absence of 'influential cases'. Homoscedasticity was evaluated by grouping participants into quartiles of the predicted scores and by applying the Levene test to the residuals. The normal distribution of the standardized residuals was investigated by conducting Kolmogorov-Smirnov tests on the residual values. The occurrence of multicollinearity was checked by calculating the Variance Inflation Factors (VIFs), which should not exceed 10 (Belsley, Kuh, & Welsch, 1980). Cook's distances were calculated to identify possible influential cases (Cook & Weisberg, 1982). As for the binary logistic regression models, multicollinearity and influential cases were checked.

To obtain normative data for the speed measures, the square root transformed CST scores are converted into standardized residuals in three steps. First, the predicted CST scores of a test subject are calculated by using the final regression models (predicted score = $B_0 + B_1X_1 + \dots + B_nX_n$). Secondly, the residuals are calculated ($e_i = - [\text{observed score} - \text{predicted score}]$). The sign of the residual score was reversed because, unlike in most tests, in the CST a higher score reflects a worse performance. Thirdly, the residuals are standardized ($Z_i = e_i / SD[\text{residual}]$). The standardized residuals can be interpreted as percentiles via a Z-distribution table with cumulative probabilities (if the model assumptions of normality and homogeneity of variance of the residuals are met), or via a table with the observed distribution of the residuals of the normative sample (if the standardized residuals are not normally distributed). More details regarding the regression-based normative method can be found elsewhere (Van Breukelen & Vlaeyen, 2005).

Likewise, normative data for the accuracy measures are established in three steps. First, the predicted probabilities of making one or more errors in the different CST Parts are calculated (i.e. $p[\text{Error CSTx}=1]$; with the symbol x reflecting Part A, B or C of the CST) using the logistic functions, i.e. $p(\text{Error CSTx}=1) = 1 / (1 + \exp[-\{B_0 + B_1X_1 + \dots + B_nX_n\}])$. Secondly, the residuals are calculated ($e_i = -[\text{observed Error} - p\{\text{Error CSTx}=1\}]$), with observed Error = 0 if the test subject made no errors in that CST Part and observed Error = 1 if one or more errors are made. Thirdly, the residuals are standardized by dividing them by the *SD* of a Bernoulli distribution with $p = p(\text{Error CSTx}=1)$, which equals $\sqrt{p[\text{Error CSTx}=1] * [1 - p\{\text{Error CSTx}=1\}]}$. The standardized residuals are then interpreted as percentiles, based on the observed distribution of the standardized residuals of the normative sample.

Results

Predictors of the CST scores

The final linear regression models for the speed measures ($\sqrt{\text{CSTa}}$, $\sqrt{\text{CSTb}}$, $\sqrt{\text{CSTc}}$ and $\sqrt{[\text{Shifting}+10]}$) are presented in Table 2. There was no serious influence of outliers (maximum Cook's distance for the final models equaled .036) or multicollinearity (maximum VIF for the final models = 1.3) observed for the models presented in Table 2, but the Levene test rejected the homogeneity of variance assumption for all the models (all $ps < .005$). The residuals were normally distributed for the $\sqrt{\text{CSTa}}$ models, but the residuals for the other models were not normally distributed (ps of the Kolmogorov-Smirnov $Zs < .005$).

Age and LE affected performance on all the CST speed measures. The influence on the predicted scores of a low LE compared with an average LE was larger than the influence of a high LE compared with an average LE (see the B 's of LE low and LE high in Table 2). For example, the effect of a low versus average LE on the $\sqrt{[\text{Shifting}+10]}$ score was comparable to the effect of about 23 years of aging, and the effect of an average versus high LE was comparable to the effect of about 14 years of aging. In addition to the linear age-effects, there were quadratic age effects for all scores with the exception of the $\sqrt{[\text{Shifting}+10]}$ score. Gender effects were only seen on $\sqrt{\text{CSTa}}$ and $\sqrt{\text{CSTb}}$, with women performing better than men. Handedness did not affect any of the CST speed scores.

The final binary logistic regression models for the CST accuracy scores are presented in Table 3. There was no serious influence of outliers (maximum Cook's distance for the final models equaled .074) or multicollinearity (maximum VIF for the final models = 1.3) observed for these models. The predicted probabilities of making errors in the CST parts B and C ($p[\text{Error CSTb}=1]$ and $p[\text{Error CSTc}=1]$) increased as a function of age and were lower for people with a higher LE. The effect of a low LE compared with a high LE on $p(\text{Error$

CSTb=1) and $p(\text{Error CSTc}=1)$ was large, and corresponded to an age effect of about 57 and 64 years respectively. The probability of making errors in the CST Part A was not related to any of the independent variables.

Table 2. Final multiple linear regression models for the CST speed measures.

Score	Variable	<i>B</i>	Std. Error <i>B</i>	<i>T</i>	Standardized <i>B</i>	<i>R</i> ²	<i>MS</i> (Error)
√ CSTa	(constant)	3.668	0.028	129.875**			
	Age	0.023	0.001	26.077**	0.525		
	Age ²	0.0002	< .001	3.880**	0.074		
	Gender	0.143	0.027	5.371**	0.101		
	LE low	0.188	0.031	6.009**	0.127		
	LE high	-0.092	0.035	-2.674	-0.055	0.381	0.310
√CSTb	(constant)	4.154	0.032	128.033**			
	Age	0.023	0.001	22.741**	0.462		
	Age ²	0.0001	< .001	2.911*	0.056		
	Gender	0.136	0.031	4.430**	0.084		
	LE low	0.365	0.036	10.154**	0.217		
	LE high	-0.146	0.040	-3.669**	-0.076	0.371	0.409
√CSTc	(constant)	5.056	0.040	125.950**			
	Age	0.034	0.001	24.112**	0.471		
	Age ²	0.0003	< .001	3.468**	0.064		
	LE low	0.576	0.050	11.446**	0.235		
	LE high	-0.308	0.055	-5.563**	-0.111	0.417	0.800
√(Shifting+10)	(constant)	4.463	0.032	139.027**			
	Age	0.018	0.001	13.707**	0.305		
	LE low	0.415	0.049	8.450**	0.202		
	LE high	-0.251	0.054	-4.675**	-0.108	0.213	0.759

Note. All scores were measured in seconds; LE = level of education. Coding of the predictors: Age = calendar age - 50; Age² = (calendar age - 50)²; Gender: Male = 1, Female = 0; LE low: low education = 1, average or high education = 0; LE high: high education = 1, low or average education = 0.

* $p < .005$; ** $p < .001$

Normative data

The CST speed measures are normed by first calculating the predicted scores (using the regression models in Table 2). Next, the residuals for each score are calculated ($e_i = \text{[observed score - predicted score]}$) and standardized ($Z_i = e_i / SD[\text{residual}]$). As described, heteroscedasticity was observed for all the models of the speed scores. This was taken into account by calculating the $SD(\text{residual})$ s per quartile of the predicted scores (Van Breukelen & Vlaeyen, 2005). Thus, the $SD(\text{residual})$ to be used in the equation that standardizes the test subject's residual score depends on the test subject's predicted score (see

Table 4). The resulting Z-values are interpreted by using a standard normal distribution table with cumulative percentages for the $\sqrt{\text{CSTa}}$ score (because the residuals were normally distributed), or by using a table with the observed distribution of the standardized residuals and their corresponding percentiles for the $\sqrt{\text{CSTb}}$, $\sqrt{\text{CSTc}}$, and $\sqrt{(\text{Shifting}+10)}$ scores (due to violations of the normality assumption; Table 5).

Table 3. Final binary logistic regression models for the CST accuracy scores (the Error scores).

Score	Variable	<i>B</i>	Std. Error <i>B</i>	Wald χ^2	<i>e</i> ^B	Nagelkerke <i>R</i> ²
Error CSTb	(constant)	-2.871	0.164	307.884**	0.057	
	Age	0.026	0.006	21.770**	1.026	
	LE low	1.253	0.194	41.670**	3.501	
	LE high	-0.234	0.297	0.622	0.791	0.135
Error CSTc	(constant)	-1.034	0.086	144.566**	0.356	
	Age	0.015	0.004	18.822**	1.015	
	LE low	0.574	0.125	20.991**	1.775	
	LE high	-0.383	0.156	6.035	0.682	0.071

Note. LE = level of education. Coding of the predictors: Age = calendar age - 50; LE low: low education = 1, average or high education = 0; LE high: high education = 1, low or average education = 0.

** $p < .001$

For example, a 40-year-old man with an average level of education took the CST. The average time to complete the CST zero trials was 5 seconds, and 20, 24 and 32 seconds to complete Part A, Part B, and Part C of the CST respectively. So, the CSTa, CSTb and CSTc scores were 15, 19, and 27, which yields a Shifting score of 10 ($= 27 - [\{15 + 19\} / 2]$). In this example we will norm the Shifting score. Thus, since this test subject's raw Shifting score was 10, the transformed Shifting score was $\sqrt{20}$ (i.e. $\sqrt{[\text{Shifting}+10]}$). The predicted $\sqrt{(\text{Shifting}+10)}$ score for this person was 4.283 ($= 4.463 + [.018 * -10] + [.415 * 0] + [-.251 * 0]$). So, the residual equals $-.189$ ($= -[\sqrt{20} - 4.283]$) and the standardized residual *Z* is $-.257$ ($= -.189 / .734$). This *Z*-value corresponds to a percentile between 20 and 50 (Table 5), which most clinicians would consider 'average' performance (see e.g. Lezak et al., 2004).

The CST accuracy scores of a test subject are normed by using the logistic regression models (Table 3). After calculation of the probabilities that a test subject makes one or more Errors ($p[\text{Error CSTx}=1]$), the residuals of the Error CSTx scores are calculated ($e_i = -[\text{observed Error} - p\{\text{Error CSTx}=1\}]$) and standardized ($Z_i = e_i / \sqrt{[p\{\text{Error CSTx}=1\} * \{1 - p(\text{Error CSTx}=1)\}]}$). These standardized values are then interpreted as percentiles by using a table that presents the observed distribution of the standardized residuals (Table 5).

For example, suppose that the 40-year-old man with an average level of education made 1 error in the CST Part C. The predicted $p(\text{Error CSTc}=1)$ would equal .234 ($= 1 / [1 + \exp \{ - (-1.034 + [.015 * -10] + [.574 * 0] + [-.384 * 0]) \}]$). The residual would equal $-.766$ ($= -[1 - .234]$), which would yield a

standardized value of - 1.809 ($= -.766 / \sqrt{.234 * \{1 - .234\}}$). This Z-value corresponds to a percentile between 5 and 10 (Table 5), which most clinicians would consider ‘borderline’ (see e.g. Lezak et al., 2004).

Table 4. Standard deviations of the residuals for the $\sqrt{\text{CSTa}}$, $\sqrt{\text{CSTb}}$, $\sqrt{\text{CSTc}}$, and $\sqrt{(\text{Shifting}+10)}$ scores.

Score	Predicted score	<i>SD</i> (residual)
$\sqrt{\text{CSTa}}$	General (all scores)	0.556
	≤ 3.479	0.462
	Between 3.480 and 3.828	0.510
	Between 3.829 and 4.236	0.564
	≥ 4.237	0.667
$\sqrt{\text{CSTb}}$	General (all scores)	0.639
	≤ 3.952	0.496
	Between 3.953 and 4.384	0.607
	Between 4.385 and 4.803	0.653
	≥ 4.804	0.770
$\sqrt{\text{CSTc}}$	General (all scores)	0.893
	≤ 4.636	0.658
	Between 4.637 and 5.274	0.789
	Between 5.275 and 5.881	0.938
	≥ 5.882	1.124
$\sqrt{(\text{Shifting}+10)}$	General (all scores)	0.870
	≤ 4.197	0.607
	Between 4.198 and 4.577	0.734
	Between 4.578 and 4.945	0.975
	≥ 4.946	1.087

Table 5. Percentiles of the standardized residuals for the $\sqrt{\text{CSTb}}$, $\sqrt{\text{CSTc}}$, $\sqrt{(\text{Shifting}+10)}$, Error CSTb and Error CSTc models.

Percentile	$\sqrt{\text{CSTb}}$	$\sqrt{\text{CSTc}}$	$\sqrt{(\text{Shifting}+10)}$	Error CSTb	Error CSTc
95	1.467	1.407	1.465	0.586	0.924
90	1.21	1.162	1.191	0.544	0.855
80	0.813	0.842	0.764	0.444	0.712
50	0.082	0.099	0.091	0.238	0.513
20	-0.768	-0.723	-0.726	0.174	-1.218
10	-1.243	-1.277	-1.221	-1.535	-1.623
5	-1.775	-1.813	-1.740	-2.236	-1.876

User-friendly normative tables

The regression-based normative method has important methodological advantages compared with a traditional normative approach, but has the drawback that the norms are not very user-friendly (Van Breukelen & Vlaeyen, 2005). Indeed, the three-step procedure to norm a test subject's scores requires quite some calculation. This may hamper the use of norms in clinical practice. Therefore, we also established normative tables so that a test subject's raw CST scores can be evaluated without calculation.

Regarding the CST speed measures, normative tables that present the raw CST scores that correspond with the percentiles 5, 10, 20, 50, 80, 90 and 95 are provided online (Tables A1 – A5). Table 6 gives an example of such a table for the Shifting score. The values in these tables were calculated by means of the three-step procedure described before. It should be emphasized that these tables present norms for the *raw* CST scores rather than for the square root transformed scores, in order to improve the user-friendliness of the norms. For example, Table 6 shows that the Shifting score of 10 that was obtained by the 40-year-old man educated to an average level from the previous example corresponds with a percentile between 20 and 50.

Regarding the accuracy measures, normative tables can be derived using two approaches. Percentiles that correspond to the dichotomized error scores (Error CST_x=0 and Error CST_x=1) can be calculated using the three-step normative procedure described above, but a drawback of this approach is that it yields normative data for the dichotomized error scores only (i.e. for the categories 'no errors' and 'one or more errors'). In other words, no distinction is made between 1, 2, 3, ... errors. This is undesirable for clinicians, because whether a patient makes 1 or 5 errors may make a large difference from a clinical perspective. The second approach to derive normative tables avoids this problem. In this approach, the final logistic models are used to determine which independent variables affect a score. Next, cumulative percentages for the observed values of the raw CST Error scores are provided for each relevant subgroup. The normative cumulative frequency tables are available online (Tables A6 – A8).

Discussion

The main aim of the present study was to evaluate whether age, gender, level of education, and/or handedness affected performance in the CST, and to determine the normal range of scores in the CST. The results showed that the linear age-effect was an important predictor of the CST speed scores ($\sqrt{\text{CSTa}}$, $\sqrt{\text{CSTb}}$, $\sqrt{\text{CSTc}}$, and $\sqrt{[\text{Shifting}+10]}$), with additional quadratic age effects on the $\sqrt{\text{CSTa}}$, $\sqrt{\text{CSTb}}$, and $\sqrt{\text{CSTc}}$ scores. Level of education also affected performance on the speed measures. We found that a low level of education compared with an average level of education had a greater impact on the predicted

Table 6. Normative table for the raw CST Shifting score stratified by age (25, 30, ..., 80 years) and level of education.

	Z	Cum.	Age in years											
	value	prob.	25	30	35	40	45	50	55	60	65	70	75	80
LE low	1.47	0.95	1.2	1.9	0.1	0.7	1.3	1.9	1.4	2.0	2.6	3.3	4.0	4.6
	1.19	0.90	2.6	3.3	1.9	2.5	3.2	3.8	3.5	4.2	4.8	5.6	6.3	7.0
	0.76	0.80	5.0	5.7	4.9	5.6	6.3	7.1	7.1	7.9	8.6	9.4	10.2	11.0
	0.09	0.50	9.0	9.8	10.4	11.2	12.1	12.9	13.7	14.6	15.5	16.4	17.3	18.3
	-0.73	0.20	14.6	15.5	18.3	19.2	20.2	21.2	23.1	24.2	25.2	26.3	27.4	28.5
	-1.22	0.10	18.3	19.3	23.6	24.7	25.7	26.8	29.6	30.8	31.9	33.1	34.3	35.5
	-1.74	0.05	22.5	23.6	29.7	30.9	32.0	33.2	37.1	38.3	39.6	40.8	42.1	43.4
LE average	Z	Cum.	Age in years											
	value	Prob.	25	30	35	40	45	50	55	60	65	70	75	80
	1.47	0.95	-0.2	0.3	0.9	0.3	0.9	1.5	2.1	0.3	0.9	1.5	2.1	1.6
	1.19	0.90	0.8	1.4	2.0	1.6	2.2	2.9	3.5	2.1	2.8	3.4	4.1	3.8
	0.76	0.80	2.6	3.2	3.9	3.9	4.5	5.2	5.9	5.2	5.9	6.6	7.4	7.4
	0.09	0.50	5.7	6.4	7.1	7.8	8.5	9.3	10.1	10.7	11.6	12.4	13.3	14.1
	-0.73	0.20	9.8	10.6	11.5	13.2	14.1	15.0	15.9	18.6	19.6	20.6	21.6	23.5
LE high	-1.22	0.10	12.6	13.5	14.3	16.8	17.8	18.7	19.7	24.0	25.1	26.2	27.3	30.1
	-1.74	0.05	15.7	16.6	17.6	20.9	21.9	22.9	24.0	30.2	31.3	32.5	33.7	37.5
	Z	Cum.	Age in years											
	Value	Prob.	25	30	35	40	45	50	55	60	65	70	75	80
	1.47	0.95	-1.7	-1.2	-0.7	-0.1	0.5	-0.2	0.4	1.0	1.6	2.2	0.5	1.0
	1.19	0.90	-0.8	-0.2	0.4	0.9	1.6	1.1	1.7	2.4	3.0	3.7	2.3	2.9
	0.76	0.80	0.9	1.5	2.1	2.7	3.4	3.3	4.0	4.7	5.4	6.1	5.3	6.1
LE high	0.09	0.50	3.7	4.4	5.1	5.8	6.5	7.2	7.9	8.7	9.5	10.3	10.9	11.7
	-0.73	0.20	7.7	8.4	9.2	10.0	10.8	12.5	13.4	14.3	15.1	16.1	18.8	19.8
	-1.22	0.10	10.3	11.1	11.9	12.8	13.7	16.1	17.0	18.0	18.9	19.9	24.3	25.3
	-1.74	0.05	13.2	14.1	15.0	15.9	16.8	20.1	21.1	22.1	23.2	24.2	30.4	31.6

Note. LE = level of education; Cum. prob. = cumulative probability.

CST scores than a high level of education compared with an average level of education. This finding is not limited to speeded tasks such as the CST but was also observed in other neuropsychological tests such as the Verbal Learning Test of Rey (Van der Elst, van Boxtel, Van Breukelen, & Jolles, 2005) and is in agreement with the reserve hypothesis of brain aging (Stern et al., 2003) and with clinical studies that found a strong association between a low level of education and dementia (De Ronchi et al., 1998). We also found gender effects on

the $\sqrt{\text{CSTa}}$ and $\sqrt{\text{CSTb}}$ scores, with females outperforming males. The absence of gender effects on the $\sqrt{\text{CSTc}}$ and $\sqrt{(\text{Shifting}+10)}$ scores may be attributable to Type II error. Indeed, because significant gender effects were found for the $\sqrt{\text{CSTa}}$ and $\sqrt{\text{CSTb}}$ scores that were also in the same direction (i.e. a positive sign of the regression weights for both scores), it was also expected that gender effects would be found on the $\sqrt{\text{CSTc}}$ or on the $\sqrt{(\text{Shifting}+10)}$ score (because Shifting is calculated as $\text{CSTc} - [\text{CSTa} + \text{CSTb}] / 2$). The effects of gender on the $\sqrt{\text{CSTc}}$ or the $\sqrt{(\text{Shifting}+10)}$ scores may however be considered to be small given the fact that they were not detected in our large sample, and thus of no relevance to take into account in the normative data. The probability of making Errors in the CST Part A was not affected by any of the demographical variables. People with a lower educational level and older people were more likely to make errors in the CST Parts B and C compared with their more highly educated and younger counterparts.

It should be noted that the raw CST Shifting score was negative for 3.2% of the normative sample. Younger, highly educated people whose scores were above the 90th percentile were more likely to have negative Shifting scores than older, less well-educated people. A possible explanation for the observation that these people completed Part C of the CST faster than the average time they needed to complete Part A and Part B is that especially younger, more highly educated people may show practice effects during the administration of the CST. However, we could not test this explanation because the order of presentation of the different CST parts was fixed in this study. Circumstantial evidence for this explanation can be found in a study of Miner and Ferraro (1998), who found that the order of presentation of the TMT Parts A and B affected the performance of 110 university graduates, suggesting that practice effects occur in young adults who are highly educated. Because Miner and Ferraro used college students, we do not know whether older or less well educated people show less practice effects, but such results have been found in studies with other cognitive tests (e.g. Dikmen, Heaton, Grant, & Temkin, 1999).

Our study had some limitations. First, our sample was not balanced in terms of level of education because there was a decrease in educational attainment as a function of age. Establishing norms that are based on an unbalanced sample is problematic when a traditional normative approach is used (see Van der Elst et al., 2006a), but this problem is minimized when a regression-based normative approach is used. Indeed, regression weights are estimated in an unbiased way no matter what the distribution of the predictors in the sample is, as long as the regression model is correctly specified, i.e. as long as no predictors or interactions or nonlinear effects are overlooked (Fox, 1997). In addition, regression analysis uses the entire sample for estimating and testing effects of predictors, thereby leading to more stable effect estimates and norms than subgroup analyses. There is only some loss of statistical power when an unbalanced sample is analyzed because the standard error of a regression weight is proportional to the

$\sqrt{\text{VIF}}$ of the predictor at hand (see e.g. Kleinbaum, Kupper, Muller & Nizam, 1998).

Secondly, the regression-based approach is more complex and less familiar to clinicians than the traditional normative approach. This is not a problem per se, but it may hamper use of regression-based norms in clinical practice. The user-friendliness of the normative data can be increased by implementing the three-step normative procedure in computer software (which can easily be done with standard software such as Microsoft Excel or with SPSS or SAS), but a disadvantage of using computerized scoring is that not all clinicians have easy access to computers. Alternatively, regression-based tables that look the same as traditional normative tables can be constructed. This is at the cost of some precision because interpolation is required when the raw test score falls between two tabulated values. With regard to the CST accuracy measures, it is at the cost of dividing the total sample into different subgroups to provide the normative data. As a result, the estimates of the population statistics are less stable, which may give rise to chance trends in the norms.

Thirdly, although several studies have suggested that the CST is a useful test for assessing concept shifting and related functions (De Groot, Hornstra, Rozendaal, & Jolles, 2003; Rasquin et al., 2004; Valentijn et al., 2005), more specific psychometric research is needed to establish the validity of the test.

Fourthly, our normative data were derived from a sample of native Dutch speakers, and it can be questioned whether these data can be applied to individuals who have a different native language or cultural background. The applicability of the CST and normative data in non-Western societies seems to be limited, because knowledge of the Western alphabet is needed to complete the test. When it comes to usage of the CST in Western societies, Møller et al. (1998) evaluated the CST performance of a control sample of people who were aged at least 60 in their study of postoperative cognitive dysfunction. This study involved people from eight European countries and the United States. The authors suggested that the CST was a culturally robust test, but more research is needed to generalize their results as the Møller et al. (1998) study involved only people aged at least 60.

In summary, the CST is a newly developed test that measures concept shifting and executive functioning. We found that females outperformed males, higher educated people outperformed lower educated people, and younger individuals outperformed older individuals. We established normative data for the CST, taking these variables into account, so that the CST can be widely used by clinicians and researchers who need to assess the executive functioning and concept shifting abilities of individuals.

Chapter 6

Assessment of information processing in working memory in applied settings: the Paper & Pencil Memory Scanning Test

Wim Van der Elst, Martin P. J. Van Boxtel, Gerard J. P. Van Breukelen & Jelle Jolles

Abstract

Background. Sternberg's Memory Scanning Test (MST; Sternberg, 1966) is a useful paradigm for evaluating the speed and efficiency of information processing in working memory. Unfortunately, the classical MST has major drawbacks for use in applied settings such as the clinic. For example, its administration time is long and the test is too difficult for old people or people with cognitive disorders. It would be advantageous to have a test for the assessment of information processing in working memory for use in applied settings, for example in differential diagnostics in clinical settings.

Method. The MST was modified into a format that makes it more appropriate for use in the clinic, the Paper & Pencil MST (P&P MST). The influence of age and age-extrinsic factors on the P&P MST was evaluated in a large sample ($N=1,839$) of healthy and cognitively intact adults (24-81 years) to establish the normal range of performance.

Results. Age and education affected all components of information processing in working memory. Gender did not affect the non-memory processing stages in the P&P MST, but affected the speed of memory scanning. An Age x Gender interaction was observed, which suggested that females who were aged below 55 scanned working memory faster than males, and vice versa for people aged above 55.

Conclusions. The established P&P MST norms provide a useful tool in applied settings when a person's memory scanning and non-memory processes in working memory are to be evaluated.

Submitted to *Psychological Medicine*

Introduction

The Memory Scanning Test (MST) was introduced by Sternberg in 1966. In the MST, a set of one to six different random digits (called the 'positive set') needs to be memorized. After a short delay, a test probe of one digit is presented after which the testee has to decide as fast as possible whether or not the test probe was a member of the positive set. This procedure is repeated multiple times with either the same or a different positive set (fixed-set procedure and varied-set procedure, respectively). The processing that occurs in the MST consists of four, supposedly additive, independent phases: (a) encoding of the stimulus, (b) working memory scanning, (c) binary decision about the nature of the response, and (d) response organization and execution (Sternberg, 1975). The MST paradigm allows for distinguishing memory (stage b) and non-memory stages (stages a, c, and d). The slope of the line that is obtained by regressing the response times on the set size provides a measure of the speed of memory scanning; and the intercept provides an estimate of the times needed to complete the non-memory stages. Due to its firm theoretical paradigm, the MST has become a popular task to study information processing in working memory in different scientific settings. Still, the use of the MST in the clinic has been very limited until now. This is, at least in part, attributable to the fact that certain test characteristics of the MST hamper routine administration in a clinical setting. First, in the generic test paradigm, a computer is needed to present the stimuli and to record the responses. Secondly, the administration time of the MST is long. Thirdly, the test is too difficult for certain clinical populations and, finally, the main outcome scores of the MST are difficult to compute.

Brand and Jolles (1987) developed a paper & pencil version of the MST (P&P MST) to overcome these problems. In this test, the participant is shown a sheet of paper that depicts a positive set of one to four letters. This sheet is replaced after five seconds by a test sheet that contains a matrix of 120 letters (for an example, see Figure 1). The participant is instructed to cross out the letters that were in the positive set. The P&P MST has several characteristics that make it a more suitable test for use in the clinic than the original MST. First, the test is in a paper & pencil format, so no computer is needed to present the stimuli or to record the responses. Paper & pencil test versions are often preferred to computer-based tests in the clinic because computers are not always available. Also, many testees are not familiar with computers (e.g. elderly and poorly educated people) which may affect their test performance adversely. Secondly, the administration of the P&P MST takes only about five minutes. Brief tests are preferred by clinicians because many people who are typically referred for cognitive evaluation (i.e. people with suspected or known brain dysfunction) are not able to undergo long test sessions. Thirdly, the P&P MST is less difficult than the original MST because the maximum size of the positive set is only four items instead of the six used in the original MST. Previous research suggests that memory span sizes greater than four may be particularly difficult for people

with cognitive disorders such as mild and moderate dementia (Storandt et al., 1984). Thus, it is important to limit the size of the positive set to four items because otherwise administration of the test would not be feasible in certain clinical populations. Fourthly, the main outcome variables of the original MST are difficult to compute. Indeed, the response times need to be regressed on the size of the positive set. The slope of this function provides a measure of memory scanning rate, and the intercept reflects the times needed to complete the operations that occur during the non-memory stages of the MST. Conducting a regression analysis is a cumbersome procedure that many clinicians may not be inclined to perform. Thus, a simplified scoring system was devised for the P&P MST to increase its user-friendliness. The time needed to complete the P&P MST Trial 1 served as the intercept measure because it reflects the response latency with a memory load of one item. The rate of working memory scanning (i.e. the slope) was quantified by calculating the average increase in the time needed to complete a test form per item extra in working memory.

Having a Sternberg paradigm-based test available for use in the clinic may be of advantage for many clinicians. For example, Ferraro and Balota (1999) found that the speed of the memory scanning (slope) and non-memory stages (intercept) of the MST were reduced in healthy older people relative to younger people. People with Alzheimer dementia showed impaired memory and non-memory stages compared with non-demented age-matched controls (Ferraro & Balota, 1999). Naus and colleagues (1977) showed similar deficits in people with Korsakoff amnesia. The memory scanning rate of non-demented Parkinson disease patients did not differ from controls, but their non-memory processing was slower (Poewe et al., 1991). Archibald and Fisk (2000) found that both relapsing-remitting and secondary-progressive multiple sclerosis patients showed a reduced working memory scanning speed relative to controls, whilst the non-memory phases in the MST were only impaired for the secondary-progressive multiple sclerosis patients. Thus, several studies suggest that the MST may be a useful test in the clinic, for example from a diagnostic viewpoint. As noted by Archibald and Fisk (2000), a problem with the original MST is that there are no normative data available for it. The lack of normative data is most likely related to the fact that using the original MST is difficult in the clinic, as discussed above. The aim of our study was to present normative data for the P&P MST, a Sternberg paradigm-based test that was especially developed for use in the clinic. In order to establish reliable normative data, the P&P MST was administered to a very large sample of cognitively intact people aged between 24 and 81 ($N=1,839$). We established norms for the intercept and slope measures of the P&P MST, and evaluated whether the 'original' and simplified procedures of calculating the slope and intercept measures yielded clinically relevant differences in the interpretation of the P&P MST performance.

Method

Participants

The data were derived from the Maastricht Aging Study (MAAS), a prospective study into the determinants of cognitive aging. The MAAS sample has been described in detail elsewhere (Jolles et al., 1995). Briefly, a large group of cognitively intact people aged between 24 and 81 underwent various cognitive tests and medical examinations. The P&P MST was administered to 1,839 people. The data of 29 participants were excluded from the analyses due to the following reasons: a score below 24 on the Mini-Mental State Examination (Folstein et al., 1975; $n=13$), the occurrence of technical problems during test administration ($n=3$), refusal or lack of motivation of the participant ($n=4$), and physical or cognitive limitations that obstructed test administration ($n=9$). Basic descriptive data for the sample are provided in Table 1. The ethnic background of all participants was Caucasian, and all participants were native Dutch speakers. Level of education (LE) was measured by classifying formal schooling into three groups – those with at most primary education (LE low), those with junior vocational training (LE average), and those with senior vocational or university training (LE high). This LE system is often used in the Netherlands (De Bie, 1987) and is comparable with the International Standard Classification of Education (UNESCO, Paris, 1976). LE low, LE average and LE high corresponded with an average of 8.59, 11.41 and 15.26 years of full-time education in the sample respectively (SD 1.93, 2.50 and 3.32 respectively).

Table 1. Descriptive characteristics of the sample that was administered the P&P MST (N=1,810).

Age		Age		Level of Education Frequency			Male:Female
Group	<i>N</i>	Mean	<i>SD</i>	Low	Average	High	Ratio
25±1 years	161	25.407	.904	21	79	61	80:81
30±1 years	150	30.592	.872	22	78	50	76:74
35±1 years	160	35.494	.986	34	76	50	78:82
40±1 years	158	40.489	.867	38	75	45	77:81
45±1 years	160	45.518	.968	50	69	41	80:80
50±1 years	157	50.286	.872	49	70	38	79:78
55±1 years	162	55.414	1.002	83	52	27	85:77
60±1 years	154	60.601	.836	70	64	19	82:72
65±1 years	159	65.433	.912	86	53	20	77:82
70±1 years	161	70.323	.817	84	57	20	87:74
75±1 years	162	75.172	.800	82	56	23	79:83
80±1 years	66	79.810	.781	41	8	15	33:33
Total	1,810	51.589	16.427	660	737	409	913:897

Note. Data on level of education were missing for 4 participants.

Procedure and Instruments

The P&P MST was administered individually at the neuropsychological laboratory of the Brain & Behaviour Institute (Maastricht, Netherlands). The P&P MST consists of four trials, in which the positive set contains 1, 2, 3, and 4 letters (i.e. “H”, “N R”, “Q W K”, and “S Z X P” respectively). The administration of each trial of the P&P MST consists of two parts. In the first part, a white sheet of paper (portrait oriented A4 format, 11.69 in x 8.26 in [29.70 cm x 20.99 cm]) that depicts the positive set in the centre of the sheet is shown to the participant for five seconds. In the second part, the sheet with the positive set is replaced by the actual test sheet. This sheet presents a matrix of 10 x 12 items printed on a sheet of white paper (portrait oriented A4 format). Twenty items are “targets” (i.e. items that were part of the positive set) and the remaining 100 items are “distractors” (i.e. items that were not part of the positive set). All items are printed in black ink in capital letters (font size 12). An example of both sheets is given in Figure 1. The participants were instructed to memorize the item(s) of the positive set in the first part. In the second part, they were asked to scan the test sheet line by line from left to right and cross out all the targets with a pencil as quickly and accurately as possible. There was no time limit to complete a sheet. The times needed to complete each test sheet were recorded (in seconds), as well as the number of omissions (i.e. the number of targets that were not crossed out) and errors (i.e. the number of distractors that were crossed out).

A practice trial was administered prior to the four test trials to make sure that the participants understood the test instructions. The positive set contained a non-letter symbol (i.e. “&”) in the practice trial. The administration order of the test trials was fixed for all participants: the trial with a positive set of 1 letter was always administered first, followed by the trials with a positive set of 2, 3 and 4 letters. These trials will be referred to as Trial 1, Trial 2, Trial 3 and Trial 4 respectively.

We used two methods to calculate the slope and intercept measures. The first method is equivalent to the method used in the original MST paradigm, and consists of regressing the response times on the set sizes. In the second method, the Trial 1 time served as the intercept score. The slope was calculated as $((\text{Trial 2 time} - \text{Trial 1 time}) + (\text{Trial 3 time} - \text{Trial 2 time}) + (\text{Trial 4 time} - \text{Trial 3 time}))/3$, which can be simplified to $([\text{Trial 4 time} - \text{Trial 1 time}]/3)$ by basic algebraic operations. The intercept and slope measures that were calculated by using the original MST method will be referred to by adding the subscript o (i.e. Intercept_o and Slope_o), the subscript s is used to indicate that the simplified calculation method was used (i.e. Intercept_s and Slope_s).

Figure 1. An example of the stimulus material for the P&P MST (Trial 3): sheet with the positive set (left) and test sheet (right). The stimulus material is presented here on a reduced scale. The true size of both sheets is 11.69 in x 8.26 in (29.70 cm x 20.99 cm).

Statistical analyses

Multiple linear regression analysis was used to determine which variables were predictive for the P&P MST scores. The raw test scores were square root transformed because preliminary data analyses showed that the residuals were positively skewed before transformation of the scores (data not shown). The P&P MST square root transformed scores were regressed on age, age^2 (which enables modelling of quadratic age associations), gender, level of education and all possible two-way interactions. To account for speed-accuracy tradeoffs, we also added the number of omissions made in Trial 1 as a predictor in the $\sqrt{\text{Intercept}_0}$ and $\sqrt{\text{Intercept}_s}$ models. The Trial 1, Trial 2, Trial 3 and Trial 4 omissions were added as extra predictors in the $\sqrt{\text{Slope}_0}$ and $\sqrt{\text{Slope}_s}$ models. These corrections were made because people who make many omissions were expected to complete the test trials faster compared with people who work more accurately and make fewer omissions. Note that there is no correction for errors required because the effects of errors are already reflected in the time scores indirectly (i.e. crossing out a distractor costs time).

The coding of the predictors to be used in the regression models was as follows. Age was centred ($\text{age} = \text{calendar age} - 50$) before computing the quadratic

terms and interactions to avoid multicollinearity (Marquardt, 1980). Gender was dummy coded with male = 1 and female = 0. Level of education (LE) was dummy coded with two dummies (LE low and LE high) and LE average as the reference category. The number of omissions made in Trial 1, Trial 2, Trial 3, and Trial 4 were dummy coded after a median split. The dummy coded omission variables are referred to by adding the subscript d. Coding was as follows: Trial 1 omissions_d = 1 if Trial 1 omissions ≥ 1 , 0 if Trial 1 omissions < 1 ; Trial 2 omissions_d = 1 if Trial 2 omissions ≥ 2 , 0 if Trial 2 omissions < 2 ; Trial 3 omissions_d = 1 if Trial 3 omissions ≥ 2 , 0 if Trial 3 omissions < 2 ; Trial 4 omissions_d = 1 if Trial 4 omissions ≥ 3 , 0 if Trial 4 omissions < 3 .

The full models (including all predictors) were then reduced stepwise by eliminating the least significant predictor if its two-tailed *p*-value was above .005. This procedure is described in detail elsewhere (Van Breukelen & Vlaeyen, 2005; Van der Elst et al., 2005). We tested the assumptions of homoscedasticity, normal distribution of the residuals, absence of multicollinearity and absence of 'influential cases'. Homoscedasticity was evaluated by grouping participants into quartiles of the predicted scores and applying the Levene test to the residuals. Normal distribution of the standardized residuals was checked by conducting Kolmogorov-Smirnov tests on the residual values. The occurrence of multicollinearity was checked by calculating the Variance Inflation Factors (VIFs), which should not exceed 10 (Belsley et al., 1980). Cook's distances were calculated to identify possible influential cases (Cook & Weisberg, 1982). Normative data are provided by converting the square root transformed P&P MST scores into standardized residuals in three steps. First, the predicted CST scores of a test subject are calculated by using the final regression models (predicted score = $B_0 + B_1X_1 + \dots + B_nX_n$). Secondly, the residuals are calculated ($e_i = - [\text{observed score} - \text{predicted score}]$). Thirdly, the residuals are standardized ($Z_i = e_i / SD[\text{residual}]$). The standardized residuals are then interpreted as percentiles via a Z-distribution table with cumulative probabilities (if the model assumptions of normality and homogeneity of variance of the residuals were met), or via a table with the observed distribution of the residuals of the normative sample (if the standardized residuals were not normally distributed). As the three-step procedure to norm a person's P&P MST scores requires quite some calculation, we also established normative tables that require no calculation at all to evaluate a person's scores.

The Bland & Altman (1986) procedure was used to evaluate to what extent the original and simplified methods of calculating the Intercept and Slope measures yielded comparable results. Thus the difference between the P&P MST standardized scores for each person ($Z(\sqrt{\text{Intercept}_o}) - Z(\sqrt{\text{Intercept}_s})$ and $Z(\sqrt{\text{Slope}_o}) - Z(\sqrt{\text{Slope}_s})$) were plotted against the mean standardized scores ($[Z(\sqrt{\text{Intercept}_o}) - Z(\sqrt{\text{Intercept}_s})] / 2$ and $[Z(\sqrt{\text{Slope}_o}) - Z(\sqrt{\text{Slope}_s})] / 2$).

All analyses were performed using the SPSS 11.5 for Windows software package. The level of α error was set to .005 for all analyses.

Results

Normative models

The final regression models for the P&P MST scores are presented in Table 2. There was no significant influence of outliers (maximum Cook's distance equalled .039) or multicollinearity (maximum VIF = 2.230) observed for these models. The Levene test rejected homogeneity of variance for all models ($p < .005$). The standard deviations of the residuals increased as a function of increasing predicted scores. Normally distributed residuals were observed for the $\sqrt{\text{Intercept}_0}$ model, but the residuals for the other models were not normally distributed (p s of the Kolmogorov-Smirnov Z s $< .005$).

Table 2. Final multiple linear regression models for the P&P MST scores.

Score	Variable	<i>B</i>	Std. Error <i>B</i>	<i>T</i>	Standardized <i>B</i>	<i>R</i> ²
$\sqrt{\text{Intercept}_0}$	(constant)	4.714	.026	183.959**		
	Age	.015	.001	17.368**	.381	
	Age ²	.0005	< .001	8.369**	.174	
	LE low	.174	.032	5.469**	.126	
	LE high	-.073	.035	-2.067	-.046	.262
$\sqrt{\text{Intercept}_5}$	(constant)	4.798	.022	221.531**		
	Age	.016	.001	21.280**	.436	
	Age ²	.0004	< .001	8.579**	.166	
	LE low	.229	.027	8.514**	.183	
	LE high	-.089	.030	-2.973*	-.062	.355
$\sqrt{\text{Slope}_0}$	(constant)	3.532	.034	105.345**		
	Age	.012	.001	8.029**	.274	
	Gender	.030	.033	.913	.021	
	LE low	.140	.039	3.623**	.096	
	LE High	-.069	.043	-1.618	-.041	
	Trial 4 omissions _d	-.116	.034	-3.468**	-.083	
	Age x Gender	-.006	.002	-3.044*	-.099	.068
$\sqrt{\text{Slope}_5}$	(constant)	3.566	.034	104.285**		
	Age	.012	.001	7.929**	.270	
	Gender	.026	.033	.778	.018	
	LE low	.146	.039	3.723**	.098	
	LE High	-.074	.043	-1.701	-.043	
	Trial 4 omissions _d	-.109	.034	-3.167*	-.075	
	Age x Gender	-.006	.002	-2.922*	-.095	.068

Note. LE = level of education; Trial 4 omissions_d: number of omissions in Trial 4 $\geq 3 = 1$, number of omissions in Trial 4 $< 3 = 0$. Coding of the predictors: Age = calendar age - 50; Age² = (calendar age - 50)²; Gender: Male = 1, Female = 0; LE low: low educational level = 1, average or high educational level = 0; LE high: high educational level = 1, low or average educational level = 0.

* $p < .005$; ** $p < .001$

The $\sqrt{\text{Intercept}_o}$ and $\sqrt{\text{Intercept}_s}$ scores were affected by Age, Age², and LE. Older and poorly educated participants were outperformed by their younger and better educated counterparts. Age, Gender, LE, and Trial 4 omissions_d affected the $\sqrt{\text{Slope}_o}$ and $\sqrt{\text{Slope}_s}$ scores. Poorly educated and older people scored higher compared with highly educated and younger people. There was a significant Age x Gender interaction: females outperformed males until the age of about 55, whilst this relationship was reversed for people older than 55 (Figure 2). The $\sqrt{\text{Slope}_o}$ and $\sqrt{\text{Slope}_s}$ scores of people who made three or more omissions in Trial 4 were lower than the scores of people who made fewer than three omissions.

The P&P MST scores are normed by first calculating the predicted scores using the regression models in Table 2. Next, the residuals for each score are calculated ($e_i = - [\text{observed score} - \text{predicted score}]$) and standardized ($Z_i = e_i / SD[\text{residual}]$). Note that a residual is usually calculated as (observed score - mean score), so without reversing the positive/negative sign, because in general a higher/lower observed score compared with the mean score signifies a better/worse performance than expected, respectively. With regard to the P&P MST test scores, however, a higher score means a worse performance. For this reason, the sign of the residual value was reversed. As heteroscedasticity was observed for all models, the $SD(\text{residual})$ s per quartile of the predicted scores should be used to standardize the residual scores (see Table 3; Van Breukelen & Vlaeyen, 2005). The resulting Z-values are interpreted by using a standard normal distribution table with cumulative percentages for the $\sqrt{\text{Intercept}_o}$ score (because the residuals of the model of this score were normally distributed), or by using Table 4 for the $\sqrt{\text{Intercept}_o}$, $\sqrt{\text{Slope}_o}$ and $\sqrt{\text{Slope}_s}$ scores (due to violations of the normality assumption).

Figure 2. Predicted $\sqrt{\text{Slope}_o}$ (left) and $\sqrt{\text{Slope}_s}$ (right) scores for people with an average educational level who made fewer than 4 omissions in Trial 4, as a function of age and gender.

Table 3. Standard deviations of the residuals for the P&P MST scores.

Score	Predicted score	<i>SD</i> (residual)
$\sqrt{\text{Intercept}}_o$	General (all scores)	.568
	≤ 4.608	.475
	Between 4.609 and 4.814	.526
	Between 4.815 and 5.167	.607
	≥ 5.168	.649
$\sqrt{\text{Intercept}}_s$	General (all scores)	.482
	≤ 4.656	.419
	Between 4.657 and 4.912	.458
	Between 4.913 and 5.266	.481
	≥ 5.267	.558
$\sqrt{\text{Slope}}_o$	General (all scores)	.678
	≤ 3.403	.594
	Between 3.404 and 3.538	.665
	Between 3.539 and 3.675	.714
	≥ 3.676	.733
$\sqrt{\text{Slope}}_s$	General (all scores)	.694
	≤ 3.433	.609
	Between 3.434 and 3.574	.672
	Between 3.575 and 3.717	.739
	≥ 3.718	.748

Table 4. Percentiles of the standardized residuals for the $\sqrt{\text{Intercept}}_s$, $\sqrt{\text{Slope}}_o$, and $\sqrt{\text{Slope}}_s$ models.

Percentile	$\sqrt{\text{Intercept}}_s$	$\sqrt{\text{Slope}}_o$	$\sqrt{\text{Slope}}_s$
95	1.480	1.500	1.437
90	1.217	1.160	1.154
75	0.705	0.693	0.682
50	0.074	0.093	0.100
25	-0.621	-0.581	-0.573
10	-1.327	-1.344	-1.319
5	-1.701	-1.771	-1.800

A comparison between the original and simplified methods of scoring the P&P MST

Figure 3 shows the Bland & Altman plots for the Intercept and Slope measures. Differences in the Z-scores as large as 0.5 between the two methods occurred frequently for the Intercept measures (26.6% of the sample) but were

rare for the Slope measures (0.4% of the sample). A difference of about 0.5 on a Z-scale may seem quite large, but it is important to note that a difference of 0.5 will only yield differences in clinical interpretations of the test scores if they are obtained by people whose mean Z-score was lower than -1.5 (see the values on the abscissas in Figure 3). For example, when percentile 5 is used as a clinical cut-off, there is agreement in 96.3% of the cases for the Intercept scores and in 98.9% for the Slope scores. These results suggest that the differences between the original and simplified methods of calculating the Intercept and Slope scores are rather small and are not likely to cause important differences in the clinical interpretation of the P&P MST performance.

Figure 3. Bland & Altman plots for the $\sqrt{\text{Intercept}}$ (left) and $\sqrt{\text{Slope}}$ (right) measures.

User-friendly normative tables

Regression-based norming has some important methodological advantages compared with a ‘traditional’ normative approach (Van Breukelen & Vlaeyen, 2005), but a disadvantage of this method is that the three-step normative method described above is cumbersome and not user-friendly. This problem can be solved by providing normative tables that require no calculation at all to evaluate a person’s P&P MST scores, but this is at the cost of some accuracy (i.e. interpolation is needed when a person’s raw test score falls between two tabulated values). Such tables are presented in the Appendix for the raw Intercept_s and Slope_s P&P MST scores (Tables A1 and A2 respectively). We provided normative tables for the Intercept_s and Slope_s measures rather than for the Intercept_o and Slope_o scores because most clinicians may be expected to use the simplified method to calculate these scores. Please note that the normative tables present norms for the *raw* Intercept_s and Slope_s scores (i.e. not for the square

root transformed scores), also in view of increasing the user-friendliness of the norms.

For example, the P&P MST was administered to a 45-year-old poorly educated woman. It took her 26s, 36s, 46s, and 60s to complete Trials 1 to 4. No omissions were made in any of the four test trials. The Intercept_s score equals 26, the Slope_s measure equals 11.3 ($[60 - 26] / 3$). A raw Intercept_s score of 26 falls into the range of percentile 25-50 (see Appendix Table A1), a raw Slope_s score of 11.3 ($= [60 - 26] / 3$) falls into the range of percentile 50-75 (see Appendix Table A2). Thus, the speed of the memory scanning and non-memory processing appears normal for this woman.

Discussion

The aim of the present study was to establish normative data for a new Sternberg paradigm-based test that was especially developed for use in the clinic, the P&P MST. We used a regression-based normative approach to derive norms for the memory scanning ($\sqrt{\text{Slope}_o}$ and $\sqrt{\text{Slope}_s}$) and non-memory ($\sqrt{\text{Intercept}_o}$ and $\sqrt{\text{Intercept}_s}$) processing phases. The results showed that both the Slope and the Intercept measures increased as a function of age, suggesting that aging affects all components of information processing involved in the P&P MST. Similar age effects were also seen in studies with the original MST (e.g. Wickens et al., 1987; Ferraro & Balota, 1999). Education also affected both the memory scanning and the non-memory processing phases involved in the P&P MST, with the effects of LE low compared with effects of LE average being larger than the effects of LE high. This result is consistent with the brain reserve hypothesis (i.e. that people with less education are more vulnerable to age-related cognitive decline and brain pathology; Stern et al., 2003). Gender did not affect the non-memory processing stages in the P&P MST, but influenced the speed of memory scanning. An Age x Gender interaction was observed, suggesting that females under 55 scanned working memory faster than males, and vice versa for people over 55 (see Figure 2). To the best of our knowledge, no other studies using a Sternberg-based task have investigated Age x Gender interactions, but Meinz and Salthouse (1998) reported that the age-related decline in other cognitive tests of measures of speed was smaller for males than females.

We used two scoring procedures to quantify the memory (Slope) and non-memory (Intercept) processing phases involved in the P&P MST. The scores that were calculated by using the original versus the simplified procedures were influenced by the same independent variables, and the regression-weights of these predictors were very similar (see *Bs* in Table 2). The clinical agreement between both scoring methods was at least 96% when percentile 5 was used as a clinical cut-off. These results justify using the simplified procedure instead of the original method. Therefore we provided normative tables for the Intercept_s

and Slope_s measures. Note that the normative table for the Slope_s measure was stratified by the Trial 4 omissions_d variable in addition to the stratification by the demographical variables (in view of the significant influence of this variable on the $\sqrt{\text{Slope}_s}$ measure; see Table 2).

This brings us to some general remarks and limitations of this study. First, the sample was Dutch-speaking and Caucasian. It is unknown to what extent language or culture affects P&P MST performance. The generalizability of our norms to non-Western populations seems limited because these people are not familiar with the Western alphabet. Møller et al. (1998) evaluated the P&P MST performance of a control sample of people who were aged at least 60 in a study on postoperative cognitive dysfunction. The results suggested that the P&P MST was a culturally robust test in a sample from eight European countries and the United States, but more research is needed to generalize these results. Secondly, the P&P MST uses letters as stimuli instead of the numbers used in the original MST. Although a digit version of the P&P MST has also been developed, we used the letter version because earlier research showed that the linear fit between response time and size of the positive set was higher for the letter version (Brand & Jolles, 1987). Thirdly, the P&P MST can be formally described as a fixed-set Sternberg procedure. There is no P&P MST that uses a varied-set procedure available. However, this is not a significant problem, as Sternberg (1975) found that the results obtained with both procedures were similar. In fact, it is reasonable to assume that a varied-set procedure is more demanding, especially for patients with cognitive disorders (Brand & Jolles, 1987), and thus too difficult for certain clinical populations (e.g. people with dementia). Fourth, in the present study we did not administer an original MST, but previous research showed that the results that were obtained with the P&P MST were equivalent to the results that were obtained with the original MST (Brand & Jolles, 1987). This suggests that both tasks measure the same underlying cognitive functions. Fifth, we used a cross-sectional design in the present study. As a consequence, the effects of age on P&P MST performance are confounded by cohort effects. However, this is not a major problem because we were primarily interested in normative trends and a cross-sectional study design is appropriate for such purposes (Hayslip & Panek, 1989). Sixth, the number of omissions made in each trial was dichotomized after a median split to be used as predictors in the regression models. It is in general not advisable to dichotomize a variable because this results in a loss of measurement information and statistical power (MacCallum et al., 2002). In the context of normative analyses, however, dichotomization is useful in view of constructing user-friendly normative tables. Indeed, if the semi-continuous raw omission variables were used as predictors in the regression models instead of the dichotomized variables, a large number of normative tables would have to be generated (i.e. a table for people who scored 0 omissions, 1 omission, 2 omissions etc.). Note that we also used the semi-continuous omission measures as predictors in preliminary data-

analyses, and that these models did not explain significantly more variance than the models that used dichotomized omission measures (data not shown).

In summary, the P&P MST distinguishes itself from the ‘original’ MST on some important points: its administration time is short, no computer is needed to administer the test, the size of the largest positive set is within the memory span capacity of people with cognitive disorders, and the scoring procedure is simple. Our aim was to evaluate the range of normal performance of the scores for this test, so that the P&P MST can be widely used by clinicians to assess the information processing in the working memory of people who are referred for cognitive evaluation.

Appendix

Table A1. Normative table for the raw P&P MST Intercepts score stratified by age and level of education.

	Z value	Cum. prob.	MALE & FEMALE											
			Age in years											
			25	30	35	40	45	50	55	60	65	70	75	80
LE low	1.48	.95	17.6	17.5	17.6	17.9	18.0	18.6	19.4	20.4	20.5	21.9	23.5	25.4
	1.22	.90	18.7	18.6	18.7	18.9	19.1	19.7	20.5	21.5	21.9	23.3	25.0	26.9
	0.71	.75	20.7	20.6	20.7	21.0	21.3	22.0	22.8	23.9	24.6	26.1	27.9	30.0
	0.07	.50	23.5	23.4	23.5	23.7	24.2	24.9	25.8	27.0	28.3	29.9	31.8	33.9
	-0.62	.25	26.6	26.5	26.6	27.0	27.6	28.4	29.3	30.5	32.5	34.3	36.3	38.6
	-1.33	.10	30.1	30.0	30.1	30.4	31.3	32.1	33.1	34.4	37.2	39.0	41.2	43.7
	-1.70	.05	32.0	31.9	32.0	32.3	33.4	34.2	35.2	36.5	39.8	41.7	43.9	46.5
LE average	1.48	.95	16.2	16.1	16.2	16.0	16.4	17.0	17.7	18.4	19.5	19.8	21.4	23.2
	1.22	.90	17.1	17.0	17.1	17.0	17.4	18.0	18.8	19.5	20.6	21.2	22.7	24.6
	0.71	.75	18.9	18.9	18.9	19.0	19.4	20.0	20.8	21.7	22.9	23.9	25.5	27.5
	0.07	.50	21.3	21.2	21.3	21.6	22.0	22.7	23.6	24.6	25.9	27.4	29.2	31.3
	-0.62	.25	24.1	24.0	24.1	24.6	25.1	25.8	26.8	28.1	29.4	31.6	33.6	35.8
	-1.33	.10	27.1	27.0	27.1	27.9	28.5	29.2	30.2	31.8	33.3	36.2	38.3	40.7
	-1.70	.05	28.7	28.6	28.7	29.8	30.3	31.1	32.1	33.8	35.4	38.8	40.9	43.4
LE high	1.48	.95	15.5	15.4	15.5	15.8	16.2	16.3	17.0	17.9	18.7	20.0	20.5	22.3
	1.22	.90	16.4	16.3	16.4	16.6	17.0	17.2	18.0	18.9	19.8	21.2	21.9	23.7
	0.71	.75	18.2	18.1	18.2	18.4	18.9	19.2	20.0	21.0	22.1	23.5	24.7	26.6
	0.07	.50	20.5	20.4	20.5	20.8	21.2	21.9	22.7	23.8	25.0	26.6	28.3	30.3
	-0.62	.25	23.2	23.1	23.2	23.5	24.0	24.9	25.8	27.0	28.5	30.1	32.6	34.8
	-1.33	.10	26.2	26.1	26.2	26.5	27.0	28.3	29.2	30.4	32.2	34.0	37.2	39.6
	-1.70	.05	27.8	27.7	27.8	28.1	28.6	30.1	31.1	32.4	34.3	36.1	39.8	42.2

Note. LE = level of education; Cum. prob. = cumulative probability.

Table A2. Normative table for the raw P&P MST Slopes score for people stratified by age, level of education, gender and Trial 4 omissions_d.

LE low	Z	Cum. value	MALE												FEMALE											
			Age in years												Age in years											
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
LE low	1.44	.95	6.4	6.5	6.7	6.8	7.0	7.1	7.3	7.4	7.6	7.7	7.9	8.1	6.4	6.3	6.6	6.4	6.7	7.0	7.3	7.6	7.9	8.3	8.6	9.0
	1.15	.90	7.5	7.6	7.8	8.0	8.2	8.3	8.4	8.6	8.8	9.0	9.1	9.3	7.3	7.3	7.6	7.5	7.8	8.2	8.5	8.8	9.2	9.5	9.9	10.3
	0.68	.75	9.5	9.7	9.9	10.1	10.3	10.4	10.6	10.8	11.0	11.2	11.4	11.6	9.0	9.1	9.4	9.5	9.9	10.3	10.6	11.0	11.4	11.8	12.3	12.7
	0.1	.50	12.3	12.6	12.8	13.0	13.2	13.4	13.6	13.9	14.1	14.3	14.5	14.8	11.2	11.6	12.0	12.4	12.8	13.2	13.7	14.1	14.6	15.0	15.5	16.0
	-0.57	.25	16.1	16.3	16.6	16.8	17.1	17.4	17.6	17.9	18.1	18.4	18.6	18.9	14.1	14.9	15.3	16.1	16.6	17.1	17.6	18.2	18.7	19.2	19.7	20.3
	-1.32	.10	20.8	21.1	21.4	21.6	21.9	22.3	22.6	22.9	23.2	23.5	23.8	24.1	17.8	19.0	19.5	20.9	21.4	22.0	22.6	23.2	23.8	24.4	25.0	25.6
	-1.80	.05	24.2	24.5	24.8	25.1	25.4	25.9	26.2	26.5	26.8	27.1	27.4	27.7	20.3	21.9	22.5	24.2	24.8	25.4	26.2	26.8	27.4	28.1	28.7	29.4
LE average	Z	Cum. value	25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
	1.44	.95	6.1	6.3	6.4	6.6	6.7	6.4	6.6	6.7	6.9	7.0	7.1	7.3	5.7	6.0	6.3	6.2	6.5	6.8	6.6	6.9	7.1	7.5	7.8	8.1
	1.15	.90	7.1	7.3	7.4	7.6	7.8	7.5	7.7	7.8	8.0	8.2	8.3	8.5	6.6	6.9	7.2	7.1	7.5	7.8	7.7	8.0	8.3	8.7	9.0	9.4
	0.68	.75	8.9	9.1	9.3	9.4	9.6	9.5	9.7	9.9	10.1	10.3	10.4	10.6	8.1	8.5	8.8	8.9	9.3	9.7	9.7	10.1	10.5	10.9	11.3	11.7
	0.1	.50	11.4	11.6	11.8	12.0	12.2	12.4	12.6	12.8	13.0	13.2	13.4	13.7	10.3	10.7	11.1	11.4	11.8	12.2	12.6	13.0	13.5	13.9	14.4	14.8
	-0.57	.25	14.6	14.9	15.1	15.3	15.6	16.1	16.4	16.6	16.9	17.1	17.4	17.6	13.1	13.5	13.9	14.7	15.1	15.6	16.4	16.9	17.4	17.9	18.4	19.0
	-1.32	.10	18.7	19.0	19.3	19.5	19.8	20.9	21.1	21.4	21.7	22.0	22.4	22.6	16.6	17.1	17.5	18.8	19.3	19.8	21.2	21.7	22.4	23.0	23.5	24.1
LE high <td>-1.80</td> <td>.05</td> <td>21.6</td> <td>21.9</td> <td>22.2</td> <td>22.5</td> <td>22.8</td> <td>24.2</td> <td>24.5</td> <td>24.8</td> <td>25.1</td> <td>25.4</td> <td>25.9</td> <td>26.2</td> <td>19.0</td> <td>19.6</td> <td>20.1</td> <td>21.7</td> <td>22.2</td> <td>22.8</td> <td>24.6</td> <td>25.2</td> <td>25.9</td> <td>26.5</td> <td>27.2</td> <td>27.8</td>	-1.80	.05	21.6	21.9	22.2	22.5	22.8	24.2	24.5	24.8	25.1	25.4	25.9	26.2	19.0	19.6	20.1	21.7	22.2	22.8	24.6	25.2	25.9	26.5	27.2	27.8
	Z	Cum. value	25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
	1.44	.95	6.2	6.4	6.5	6.2	6.4	6.5	6.7	6.3	6.5	6.6	6.8	6.9	5.4	5.6	5.9	6.2	6.5	6.4	6.7	6.5	6.8	7.1	7.4	7.7
	1.15	.90	7.1	7.3	7.4	7.2	7.4	7.5	7.7	7.4	7.6	7.8	7.9	8.1	6.2	6.5	6.8	7.1	7.4	7.4	7.7	7.6	7.9	8.2	8.6	8.9
	0.68	.75	8.7	8.9	9.1	9.0	9.2	9.4	9.5	9.4	9.6	9.8	10.0	10.2	7.7	8.0	8.4	8.7	9.1	9.2	9.6	9.7	10.0	10.4	10.8	11.2
	0.1	.50	10.9	11.1	11.3	11.5	11.7	11.9	12.1	12.3	12.5	12.7	12.9	13.1	9.8	10.2	10.6	11.0	11.4	11.7	12.1	12.5	12.9	13.4	13.8	14.3
	-0.57	.25	13.8	14.0	14.3	14.8	15.0	15.2	15.5	16.0	16.3	16.5	16.7	17.0	12.5	13.0	13.4	13.8	14.3	15.0	15.5	16.3	16.8	17.3	17.8	18.3
Trial 4 omissions > 3	-1.32	.10	17.4	17.7	17.9	18.9	19.1	19.4	19.7	20.7	21.0	21.3	21.6	21.8	16.0	16.4	16.9	17.4	17.9	19.2	19.7	21.0	21.6	22.3	22.8	23.4
	-1.80	.05	19.9	20.2	20.5	21.8	22.1	22.4	22.6	24.1	24.4	24.7	25.0	25.3	18.4	18.9	19.4	20.0	20.5	22.1	22.7	24.4	25.0	25.8	26.4	27.0

Trial & omissions %	Z value	Cum. prob.	MALE												FEMALE											
			Age in years												Age in years											
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80
LE low		.95	6.3	6.5	6.6	6.8	6.4	6.6	6.7	6.9	7.0	7.2	7.3	7.5	5.9	6.2	6.5	6.3	6.6	6.5	6.8	7.0	7.3	7.7	8.0	8.3
	1.44	.90	7.3	7.5	7.6	7.8	7.5	7.7	7.9	8.0	8.2	8.3	8.5	8.7	6.8	7.1	7.4	7.3	7.7	7.6	7.9	8.2	8.5	8.9	9.2	9.6
	0.68	.75	9.1	9.3	9.5	9.7	9.6	9.8	10.0	10.1	10.3	10.5	10.7	10.9	8.3	8.7	9.0	9.1	9.5	9.6	10.0	10.3	10.7	11.1	11.5	11.9
	0.10	.50	11.6	11.8	12.1	12.3	12.4	12.6	12.9	13.1	13.3	13.5	13.7	13.9	10.5	10.9	11.3	11.7	12.1	12.5	12.9	13.3	13.8	14.2	14.7	15.1
	-0.57	.25	14.9	15.2	15.4	15.6	16.2	16.4	16.7	16.9	17.2	17.5	17.7	18.0	13.3	13.8	14.2	15.0	15.4	16.2	16.7	17.2	17.7	18.2	18.8	19.3
	-1.32	.10	19.1	19.3	19.6	19.9	20.9	21.2	21.5	21.8	22.1	22.4	22.7	23.0	16.9	17.4	17.9	19.1	19.6	21.0	21.5	22.2	22.7	23.3	23.9	24.5
	-1.80	.05	22.0	22.3	22.5	22.8	24.3	24.6	24.9	25.2	25.7	26.0	26.3	26.6	19.4	19.9	20.4	22.0	22.6	24.3	24.9	25.7	26.3	26.9	27.6	28.2
LE average		.95	6.0	6.2	6.3	6.5	6.2	6.3	6.5	6.6	6.8	6.5	6.6	6.8	5.2	5.5	5.8	6.1	6.4	6.2	6.5	6.3	6.6	6.9	7.2	7.5
	1.15	.90	6.9	7.1	7.2	7.4	7.2	7.3	7.5	7.7	7.8	7.6	7.7	7.9	6.0	6.3	6.6	6.9	7.3	7.2	7.5	7.4	7.8	8.1	8.4	8.7
	0.68	.75	8.5	8.7	8.9	9.0	9.0	9.1	9.3	9.5	9.7	9.6	9.8	10.0	7.5	7.8	8.2	8.5	8.9	9.0	9.4	9.4	9.8	10.2	10.5	10.9
	0.10	.50	10.7	10.9	11.1	11.3	11.5	11.7	11.9	12.1	12.3	12.5	12.7	12.9	9.6	10.0	10.3	10.7	11.1	11.5	11.9	12.3	12.7	13.1	13.6	14.0
	-0.57	.25	13.6	13.8	14.0	14.2	14.7	15.0	15.2	15.4	15.7	16.2	16.5	16.7	12.3	12.7	13.1	13.6	14.0	14.8	15.2	16.0	16.5	17.0	17.5	18.0
	-1.32	.10	17.1	17.4	17.6	17.9	18.8	19.1	19.4	19.6	19.9	21.0	21.2	21.5	15.7	16.2	16.6	17.1	17.6	18.9	19.4	20.7	21.3	21.8	22.5	23.1
	-1.80	.05	19.6	19.9	20.2	20.4	21.7	22.0	22.3	22.6	22.9	24.3	24.6	24.9	18.1	18.6	19.1	19.7	20.2	21.8	22.3	24.1	24.7	25.3	26.0	26.7
LE high		.95	5.7	5.8	6.0	6.1	6.3	6.4	6.1	6.3	6.4	6.6	6.7	6.4	4.9	5.1	5.4	5.7	6.0	6.3	6.1	6.4	6.7	6.6	6.9	7.1
	1.15	.90	6.5	6.7	6.8	7.0	7.2	7.3	7.1	7.3	7.4	7.6	7.7	7.5	5.7	6.0	6.3	6.6	6.9	7.2	7.1	7.4	7.8	7.7	8.0	8.3
	0.68	.75	8.1	8.3	8.4	8.6	8.8	9.0	8.9	9.1	9.2	9.4	9.6	9.5	7.1	7.4	7.8	8.1	8.5	8.8	8.9	9.3	9.6	9.7	10.1	10.5
	0.10	.50	10.2	10.4	10.6	10.8	11.0	11.2	11.4	11.6	11.8	12.0	12.2	12.4	9.1	9.5	9.9	10.3	10.6	11.0	11.4	11.8	12.2	12.6	13.0	13.5
	-0.57	.25	13.0	13.2	13.5	13.7	13.9	14.1	14.6	14.9	15.1	15.3	15.6	16.1	11.8	12.2	12.6	13.0	13.5	13.9	14.7	15.1	15.6	16.4	16.9	17.4
	-1.32	.10	16.5	16.7	17.0	17.2	17.5	17.7	18.7	19.0	19.2	19.5	19.8	20.8	15.1	15.6	16.1	16.5	17.0	17.5	18.7	19.3	19.8	21.1	21.7	22.4
	-1.80	.05	19.0	19.2	19.5	19.8	20.0	20.3	21.6	21.9	22.2	22.5	22.7	24.2	17.5	18.0	18.5	19.0	19.5	20.1	21.6	22.2	22.8	24.5	25.1	25.9

Note. LE = level of education, Cum. prob. = cumulative probability.

Chapter 7

The Letter Digit Substitution Test: normative data for 1,858 healthy participants aged 24-81 from the Maastricht Aging Study (MAAS): influence of age, education and sex

Wim Van der Elst, Martin P. J. Van Boxtel, Gerard J. P. Van Breukelen & Jelle Jolles

Abstract

The Letter Digit Substitution Test (LDST) is based on earlier developed substitution tests (e.g. Digit Symbol Substitution Test; Wechsler, 1955, 1981) but uses over-learned signs instead of the symbols used in other substitution tests.

The written and oral versions of the LDST were administered to a large, cognitively screened sample ($N=1,858$) of adults aged 24 to 81 years. Age was the most important predictor of LDST performance, and females outperformed males. A low level of education profoundly influenced LDST performance: the effect of a low versus high level of education on LDST performance was comparable to about 20 years of aging. Regression-based normative data were prepared for both the written and oral versions of the LDST.

Published in the *Journal of Clinical and Experimental Neuropsychology*. 2006; 28, 998-1009

Introduction

Substitution tests are widely used as clinical and research tools in neuropsychology (Lezak, 1995), the best known of which is the Digit Symbol Substitution Test, one of the subtests from the Wechsler Intelligence Scales (Wechsler, 1955, 1981). Several adaptations of the test now exist (e.g. Smith, 1982). Substitution tests are essentially speed-dependent tasks that require the participant to match particular signs – symbols, digits, or letters – to other signs within a specified time period. Substitution tests are sensitive to brain dysfunction (Lezak, 1995; Spreen & Strauss, 1998) in a non-specific way because their performance draws on many different processes: the simple responses generated in substitution tests depend on the integration of complex neuropsychological processes, including visual scanning, mental flexibility, sustained attention, psychomotor speed, and speed of information processing (Lezak, 1995; Van Hoof, Jogems-Kosterman, Sabbe, Zitman, & Hulstijn, 1998). This non-specific sensitivity to brain dysfunction, combined with the possibility of group administration and the short test time, makes substitution tests highly suitable as screening instruments. However, adequate screening requires adequate normative data. The main objective of this study was to derive normative data for a newly developed substitution test, the Letter Digit Substitution Test (LDST; Jolles, Houx, Van Boxtel, & Ponds, 1995).

The LDST is an adaptation of earlier substitution tests, notably the Digit Symbol Substitution Test (DSST; Wechsler, 1955, 1981) and the Symbol Digit Modalities Test (SDMT; Smith, 1982). In the DSST, digits are presented and the test participant has to respond by writing the corresponding symbol in a blank space, according to a key, whilst this relation is reversed in the SDMT. The SDMT has an advantage over the DSST because the response can be given in writing or verbally, which makes it possible to compare a participant's written and spoken performance (Lezak, 1995). The LDST differs from the DSST and the SDMT in that the key consists of 'over-learned' signs, i.e. letters and digits. Thus subjects only have to learn the letter-digit association, whereas in other substitution tests subjects also have to learn abstract symbols. This makes performance relatively more dependent on memory and complex visual processes than it is in the LDST. Thus, the rationale behind the LDST's development was to provide a more specific measure of information processing speed.

As previous research suggested that demographic factors affect LDST performance (Van Boxtel et al., 1997), we investigated to what extent age, sex, and education influence LDST performance, in order to correct the normative data for these variables.

Method

Participants

Data used for this study were derived from the Maastricht Aging Study (MAAS), a prospective study of the determinants of cognitive aging (Bosma, Van Boxtel, Ponds, Houx, & Jolles, 2003; Jolles et al., 1995; Van Boxtel et al., 2000). Participants were recruited from the Registration Network Family Practices (*RegistratieNet Huisartspraktijken*, RNH), which includes 80,000 people who live in the province of Limburg (the Netherlands). The RNH classifies health problems according to the International Classification of Health Problems in Primary Care (Lamberts & Wood, 1987). The use of the RNH as a sample frame rather than a general population sample enables the selection of an eligible study sample beforehand. Thus individuals with medical conditions known to interfere with cognition, such as cerebrovascular pathology, tumors of the nervous system, multiple sclerosis, epilepsy, parkinsonism, dementia, organic psychosis, schizophrenia, affective psychosis, and mental retardation, could be excluded from the sample frame. In total 10,396 individuals aged between 24 and 81 years were randomly drawn from the RNH. They were informed of the study by their general practitioner, which was expected to have a facilitative effect on compliance. They were asked to return a prepaid postcard indicating their willingness, or not, to participate in the study. In total 3,531 individuals refused to participate in the study (33.96%), 4,490 agreed to participate (43.19%), and 2,375 individuals did not return the postcard (22.85%). Potential participants were then screened in a semi-structured interview to check for additional exclusion criteria that were not coded in the RNH (i.e. a history of transient ischemic attacks, brain surgery, hemodialysis for renal failure, electroconvulsive therapy, and, chronic psychotropic drug use), which led to the exclusion of 301 individuals. Of the remaining 4,189 participants, 1,858 were randomly selected from 12 age categories (25 ± 1 years, 30 ± 1 years, ..., $80 \pm$ years) for participation in the study.

Not all data for the 1,858 participants administered the LDST were analyzed. Thirteen participants scored below 24 on the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975) and were therefore excluded, since they could not be considered to be 'cognitively intact'. Basic descriptive data for the sample are provided in Table 1. Level of education (LE) was assessed by classifying formal schooling using a Dutch system (De Bie, 1987) that is comparable with the International Standard Classification of Education (UNESCO, Paris, 1976). Three educational groups were formed: primary education (LE low), junior vocational training (LE average), and senior vocational or academic training (LE high), corresponding to an average of 8.56, 11.37, and 15.23 years of full-time education (SD 1.92, 2.50, and 3.32, respectively). The Table also gives the average IQ (and SD) per Age group, measured with the shortened Groningen Intelligence Test (GIT; Luteijn & Van der Ploeg, 1983), a Dutch test

that is comparable with the Wechsler Adult Intelligence Scale (Wechsler, 1981). The occupational status of the sample was as follows: full-time student (0.98%; $n=18$), paid employment (49.38%; $n=911$), unpaid employment (voluntary work or running the household; 24.50%; $n=452$), unemployed (2.33%; $n=43$), retired (14.96%; $n=276$), sick leave/disabled (4.44%; $n=82$), and unknown (3.41%; $n=63$). All participants were Caucasian and native Dutch speakers.

Table 1. Descriptive characteristics of the sample ($N=1,845$).

Age Group	N	Age		LE Frequency			Male:Female Ratio	GIT IQ	
		Mean	SD	Low	Average	High		Mean	SD
25±1 year	160	25.41	.90	21	78	61	79:81	111.76	10.03
30±1 year	153	30.61	.88	21	79	53	79:74	112.21	8.53
35±1 year	160	35.54	.91	34	77	49	80:80	113.66	11.35
40±1 year	159	40.50	.87	38	75	46	78:81	112.87	11.19
45±1 year	164	45.54	.92	52	72	40	81:83	112.35	13.62
50±1 year	163	50.30	.90	53	72	38	81:82	114.12	14.47
55±1 year	163	55.45	.90	85	51	27	83:80	113.33	14.56
60±1 year	161	60.61	.84	75	66	19	84:77	112.61	13.93
65±1 year	161	65.46	.87	89	52	20	80:81	111.39	12.54
70±1 year	163	70.33	.83	86	58	19	85:78	116.18	12.48
75±1 year	166	75.19	.81	84	57	24	83:83	118.41	13.18
80±1 year	72	79.77	.83	46	10	15	35:37	118.44	12.40
Total	1,845	51.80	16.48	684	747	411	928:917	113.75	12.66

Note. LE = Level of Education, GIT = Groningen Intelligence Test; Data on level of education were missing from 3 participants.

Procedure and instruments

The LDST test sheet is given in the Appendix (Figure 1). The key gives the numbers 1 to 9, each paired with a different letter (see also Smith, 1982; Wechsler, 1955, 1981); the test items are printed beneath the key. Participants are required to replace the randomized letters with the appropriate digit indicated by the key. The first 10 items are used as practice items, to ensure that participants understand the test instructions. After completion of these items, participants are instructed to replace the remaining items as quickly as possible. The key and the stimuli are the same for the oral and written versions of the LDST. The number of correct substitutions made in 60 seconds is the dependent variable for both test versions. The written LDST version was administered first, immediately followed by the oral version. Tests were administered by five trained assistants at the neuropsychological test laboratory of the University Hospital Maastricht (the Netherlands).

Data analysis

The normative procedure for the oral and written LDST scores involved the fitting of multiple linear regression models adjusted for Age (in years), Sex, and Level of Education. The level of α error was set to .005. Level of education (LE) was dummy coded with LE average as the reference category. The core assumptions of regression analysis (homoscedasticity, normal distribution of the residuals, absence of multicollinearity, and absence of 'influential cases') were tested for each model. Homoscedasticity was evaluated by visual inspection of the scatter plots of the residuals on the predicted values. The normal distribution of the residuals was investigated by visual inspection of the histograms and the normal probability plots. The occurrence of multicollinearity was checked by calculating the Variance Inflation Factors (VIFs), which should not exceed 10 (Belsley, Kuh, & Welsch, 1980). Cook's distances were computed to identify possible influential cases.

Normative data can then be obtained by calculating the residuals for the LDST scores ($e_i = \text{observed score} - \text{predicted score}$). The residuals are then standardized ($Z_i = e_i / \text{SD}[\text{residual}]$) so that the performance of the participant can be evaluated via a Z distribution table with cumulative probabilities.

All analyses were performed using the SPSS 10.0 for Macintosh software package.

Results

Linear multiple regression models were fitted for the oral and written LDST scores. The residuals were sufficiently normally distributed, and no heteroscedasticity was observed. VIFs of the predictors in the regression models had a maximum value of 1.30, which is well below the cut-off value of 10. The outliers had virtually no effect (maximum Cook's distance .02).

Table 2 presents the regression models. Age and LE low had a significant negative ($p < .001$) influence on the predicted LDST scores, whereas LE high had a significant positive ($p < .005$) influence. The predicted LDST scores of women were significantly higher ($p < .005$) than those of men. LE low had a greater effect on the predicted test scores than did LE high (see standardized B's, Table 2), which indicated that the predicted difference in performance between the LE low and LE average groups was greater than the difference between the LE high and LE average groups.

Combining these regression models with the standard deviations of the residuals provides normative data. First, the predicted values of the scores (predicted y_i) for the written and oral versions of the LDST are calculated by inserting the coded values of the predictor variables in the regression models (Table 2). Next, the residuals of both scores are calculated ($e_i = \text{observed } y_i - \text{predicted } y_i$) and then standardized ($Z_i = e_i / \text{SD}(\text{residual})$). The SD(residual) equals 5.63

for the written LDST scores and 6.60 for the oral LDST scores. An individual's performance can then be determined by using a standard *Z* distribution table with corresponding cumulative probabilities.

Table 2. Multiple linear regression models of the written and oral LDST scores with Age, Sex, LE low and LE high as predictors.

LDST version	Variable	B	Std. Error B	t	Standardized B	R ²	SD(residual)
Written	(constant)	48.27	.48	101.68**			
	Age	-.28	.01	-32.75**	-.56		
	Sex	-.81	.27	-3.04*	-.05		
	LE Low	-4.53	.31	-14.61**	-.27		
	LE High	1.12	.35	3.20**	.06	.52	5.63
Oral	(constant)	53.20	.56	95.52**			
	Age	-.26	.01	-26.00**	-.50		
	Sex	-1.36	.31	-4.36**	-.08		
	LE low	-3.70	.36	-10.17**	-.21		
	LE high	1.24	.41	3.03*	.06	.39	6.60

Note. LE = Level of Education. Coding of the predictors: Age = age in years; Sex: Male = 1, Female = 0; LE Low: Low Education = 1, Average or High Education = 0; LE High: High Education = 1, Low or Average Education = 0.

* $p < .005$; ** $p < .001$

Tables 3 and 4 provide normative LDST data based on the regression models in Table 2 stratified by age, level of education, and sex. If an individual is not exactly 25, 30, ..., 80 years old, then the person's age should be rounded up to the closest age given. If the performance of such an individual is borderline-normal according to these tables (e.g. *Z* value ≈ -1.28), the regression models presented in Table 2 can be used to determine the exact *Z* values.

Discussion

Age was the most important predictor of LDST performance, with a significant additional influence of education and sex. Women and higher educated participants outperformed men and lower educated participants, respectively. The predicted LDST scores of the low-educated participants were about 5 points lower than the predicted LDST scores of the high-educated participants (for the written version 5.65 points; for the oral version 4.94 points). This effect of education (high versus low) was equivalent to the effect of about 20 years of aging (for the written version 20.18 years; for the oral version 19.00 years). For example, the predicted LDST scores of a 70-year-old high-educated person were equal to the predicted scores of a 50-year-old low-educated person of the same sex.

Table 3. Normative data for the written LDST version stratified by age (25, 30, ..., 80 years), sex, and level of education. The raw test score leading to a particular Z-value is given for Z-values indicating the percentiles 1, 5, 10, 20, 50, 80, 90, 95 and 99.

	Z	Cum. prob.	Male															Female														
			Age in years															Age in years														
			25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80						
LE LOW	-2.33	.01	22.9	21.5	20.1	18.7	17.4	16.0	14.6	13.2	11.8	10.4	9.0	7.7	23.7	22.3	20.9	19.6	18.2	16.8	15.4	14.0	12.6	11.2	9.9	8.5						
	-1.64	.05	26.8	25.4	24.0	22.6	21.2	19.9	18.5	17.1	15.7	14.3	12.9	11.5	27.6	26.2	24.8	23.4	22.0	20.7	19.3	17.9	16.5	15.1	13.7	12.4						
	-1.28	.10	28.8	27.4	26.0	24.7	23.3	21.9	20.5	19.1	17.7	16.3	15.0	13.6	29.6	28.2	26.8	25.5	24.1	22.7	21.3	19.9	18.5	17.1	15.8	14.4						
	-0.84	.20	31.3	29.9	28.5	27.1	25.7	24.4	23.0	21.6	20.2	18.8	17.4	16.0	32.1	30.7	29.3	27.9	26.5	25.2	23.8	22.4	21.0	19.6	18.2	16.9						
	0	.50	36.0	34.6	33.2	31.9	30.5	29.1	27.7	26.3	24.9	23.5	22.2	20.8	36.8	35.4	34.0	32.7	31.3	29.9	28.5	27.1	25.7	24.3	23.0	21.6						
	0.84	.80	40.7	39.3	38.0	36.6	35.2	33.8	32.4	31.0	29.7	28.3	26.9	25.5	41.5	40.2	38.8	37.4	36.0	34.6	33.2	31.8	30.5	29.1	27.7	26.3						
	1.28	.90	43.2	41.8	40.4	39.1	37.7	36.3	34.9	33.5	32.1	30.7	29.4	28.0	44.0	42.6	41.2	39.9	38.5	37.1	35.7	34.3	32.9	31.5	30.2	28.8						
	1.64	.95	45.2	43.8	42.5	41.1	39.7	38.3	36.9	35.5	34.2	32.8	31.4	30.0	46.0	44.7	43.3	41.9	40.5	39.1	37.7	36.3	35.0	33.6	32.2	30.8						
	2.33	.99	49.1	47.7	46.3	45.0	43.6	42.2	40.8	39.4	38.0	36.6	35.3	33.9	49.9	48.5	47.1	45.8	44.4	43.0	41.6	40.2	38.8	37.5	36.1	34.7						
	Z	Cum. prob.	25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80						
LE AVERAGE	-2.33	.01	27.4	26.0	24.7	23.3	21.9	20.5	19.1	17.7	16.3	15.0	13.6	12.2	28.2	26.8	25.5	24.1	22.7	21.3	19.9	18.5	17.2	15.8	14.4	13.0						
	-1.64	.05	31.3	29.9	28.5	27.2	25.8	24.4	23.0	21.6	20.2	18.8	17.5	16.1	32.1	30.7	29.3	28.0	26.6	25.2	23.8	22.4	21.0	19.7	18.3	16.9						
	-1.28	.10	33.3	32.0	30.6	29.2	27.8	26.4	25.0	23.6	22.3	20.9	19.5	18.1	34.1	32.8	31.4	30.0	28.6	27.2	25.8	24.4	23.1	21.7	20.3	18.9						
	-0.84	.20	35.8	34.4	33.0	31.7	30.3	28.9	27.5	26.1	24.7	23.3	22.0	20.6	36.6	35.2	33.8	32.5	31.1	29.7	28.3	26.9	25.5	24.2	22.8	21.4						
	0	.50	40.5	39.2	37.8	36.4	35.0	33.6	32.2	30.8	29.5	28.1	26.7	25.3	41.3	40.0	38.6	37.2	35.8	34.4	33.0	31.6	30.3	28.9	27.5	26.1						
	0.84	.80	45.3	43.9	42.5	41.1	39.7	38.3	37.0	35.6	34.2	32.8	31.4	30.0	46.1	44.7	43.3	41.9	40.5	39.1	37.8	36.4	35.0	33.6	32.2	30.8						
	1.28	.90	47.7	46.4	45.0	43.6	42.2	40.8	39.4	38.0	36.7	35.3	33.9	32.5	48.5	47.2	45.8	44.4	43.0	41.6	40.2	38.8	37.5	36.1	34.7	33.3						
	1.64	.95	49.8	48.4	47.0	45.6	44.2	42.8	41.5	40.1	38.7	37.3	35.9	34.5	50.6	49.2	47.8	46.4	45.0	43.6	42.3	40.9	39.5	38.1	36.7	35.3						
	2.33	.99	53.6	52.3	50.9	49.5	48.1	46.7	45.3	43.9	42.6	41.2	39.8	38.4	54.4	53.1	51.7	50.3	48.9	47.5	46.1	44.8	43.4	42.0	40.6	39.2						
	Z	Cum. prob.	25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80						
LE HIGH	-2.33	.01	28.5	27.2	25.8	24.4	23.0	21.6	20.2	18.8	17.5	16.1	14.7	13.3	29.4	28.0	26.6	25.2	23.8	22.4	21.0	19.7	18.3	16.9	15.5	14.1						
	-1.64	.05	32.4	31.0	29.7	28.3	26.9	25.5	24.1	22.7	21.3	20.0	18.6	17.2	33.2	31.8	30.5	29.1	27.7	26.3	24.9	23.5	22.2	20.8	19.4	18.0						
	-1.28	.10	34.5	33.1	31.7	30.3	28.9	27.5	26.1	24.8	23.4	22.0	20.6	19.2	35.3	33.9	32.5	31.1	29.7	28.3	26.9	25.6	24.2	22.8	21.4	20.0						
	-0.84	.20	36.9	35.5	34.2	32.8	31.4	30.0	28.6	27.2	25.8	24.5	23.1	21.7	37.7	36.3	35.0	33.6	32.2	30.8	29.4	28.0	26.7	25.3	23.9	22.5						
	0	.50	41.7	40.3	38.9	37.5	36.1	34.7	33.3	32.0	30.6	29.2	27.8	26.4	42.5	41.1	39.7	38.3	36.9	35.5	34.1	32.8	31.4	30.0	28.6	27.2						
	0.84	.80	46.4	45.0	43.6	42.2	40.8	39.5	38.1	36.7	35.3	33.9	32.5	31.1	47.2	45.8	44.4	43.0	41.6	40.3	38.9	37.5	36.1	34.7	33.3	31.9						
	1.28	.90	48.9	47.5	46.1	44.7	43.3	41.9	40.5	39.2	37.8	36.4	35.0	33.6	49.7	48.3	46.9	45.5	44.1	42.7	41.3	40.0	38.6	37.2	35.8	34.4						
	1.64	.95	50.9	49.5	48.1	46.7	45.3	44.0	42.6	41.2	39.8	38.4	37.0	35.6	51.7	50.3	48.9	47.5	46.1	44.8	43.4	42.0	40.6	39.2	37.8	36.4						
	2.33	.99	54.8	53.4	52.0	50.6	49.2	47.8	46.4	45.1	43.7	42.3	40.9	39.5	55.6	54.2	52.8	51.4	50.0	48.6	47.3	45.9	44.5	43.1	41.7	40.3						
	Z	Cum. prob.	25	30	35	40	45	50	55	60	65	70	75	80	25	30	35	40	45	50	55	60	65	70	75	80						

Note. LE = Level of Education.

Table 4. Normative data for the oral LDST version stratified by age (25, 30, ..., 80 years), sex, and level of education. The raw test score leading to a particular Z-value is given for Z-values indicating the percentiles 1, 5, 10, 20, 50, 80, 90, 95 and 99.

Z	Cum. prob.	Male															Female																			
		Age in years															Age in years																			
		25	30	35	40	45	50	55	60	65	70	75	80	80	75	70	65	60	55	50	45	40	35	30	25											
LE LOW	-2.33	.01	26.3	25.0	23.7	22.5	21.2	19.9	18.6	17.3	16.0	14.7	13.4	12.1	10.8	9.5	8.2	7.0	5.8	4.5	3.2	2.0	0.8	27.7	26.4	25.1	23.8	22.5	21.2	19.9	18.7	17.4	16.1	14.8	13.5	
	-1.64	.05	30.9	29.6	28.3	27.0	25.7	24.4	23.1	21.8	20.6	19.3	18.0	16.7	15.4	14.1	12.8	11.5	10.2	8.9	7.6	6.3	5.0	34.6	33.3	32.0	30.7	29.4	28.2	26.9	25.6	24.3	23.0	21.7	20.4	
	-1.28	.10	33.2	32.0	30.7	29.4	28.1	26.8	25.5	24.2	22.9	21.6	20.3	19.1	17.8	16.5	15.2	13.9	12.6	11.3	10.0	8.7	37.5	36.2	34.9	33.6	32.4	31.1	29.8	28.5	27.2	25.9	24.6	23.3		
	-0.84	.20	36.2	34.9	33.6	32.3	31.0	29.7	28.4	27.1	25.8	24.5	23.3	22.0	20.7	19.4	18.1	16.8	15.5	14.2	12.9	11.6	10.3	40.8	39.5	38.2	36.9	35.6	34.3	33.0	31.7	30.4	29.1	27.8	26.5	
	0	.50	41.7	40.4	39.1	37.8	36.5	35.2	34.0	32.7	31.4	30.1	28.8	27.5	26.2	24.9	23.6	22.3	21.0	19.7	18.4	17.1	15.8	44.1	42.8	41.5	40.2	38.9	37.6	36.3	35.0	33.7	32.4	31.1	30.0	28.9
LE AVERAGE	0.84	.80	47.2	45.9	44.7	43.4	42.1	40.8	39.5	38.2	36.9	35.6	34.3	33.0	31.7	30.4	29.1	27.8	26.5	25.2	23.9	22.6	21.3	48.6	47.3	46.0	44.7	43.4	42.1	40.9	39.6	38.3	37.0	35.7	34.4	
	1.28	.90	50.1	48.8	47.6	46.3	45.0	43.7	42.4	41.1	39.8	38.5	37.2	35.9	34.6	33.3	32.0	30.7	29.4	28.1	26.8	25.5	24.2	51.5	50.2	48.9	47.6	46.3	45.0	43.8	42.5	41.2	39.9	38.6	37.3	
	1.64	.95	52.5	51.2	49.9	48.6	47.4	46.1	44.8	43.5	42.2	40.9	39.6	38.3	37.0	35.7	34.4	33.1	31.8	30.5	29.2	27.9	26.6	53.9	52.6	51.3	50.0	48.7	47.4	46.1	44.8	43.6	42.3	41.0	39.7	
	2.33	.99	57.1	55.8	54.5	53.2	51.9	50.6	49.3	48.0	46.7	45.5	44.2	42.9	41.6	40.3	39.0	37.7	36.4	35.1	33.8	32.5	31.2	58.4	57.1	55.8	54.6	53.3	52.0	50.7	49.4	48.1	46.8	45.5	44.2	
	Z	value	25	30	35	40	45	50	55	60	65	70	75	80	80	75	70	65	60	55	50	45	40	35	30	25	30	35	40	45	50	55	60	65	70	75
LE HIGH	-2.33	.01	30.0	28.7	27.4	26.2	24.9	23.6	22.3	21.0	19.7	18.4	17.1	15.8	14.5	13.2	11.9	10.6	9.3	8.0	6.7	5.4	4.1	31.4	30.1	28.8	27.5	26.2	24.9	23.6	22.4	21.1	19.8	18.5	17.2	
	-1.64	.05	34.6	33.3	32.0	30.7	29.4	28.1	26.8	25.5	24.3	23.0	21.7	20.4	19.1	17.8	16.5	15.2	13.9	12.6	11.3	10.0	8.7	35.9	34.6	33.4	32.1	30.8	29.5	28.2	26.9	25.6	24.3	23.0	21.7	
	-1.28	.10	37.0	35.7	34.4	33.1	31.8	30.5	29.2	27.9	26.6	25.3	24.1	22.8	21.5	20.2	18.9	17.6	16.3	15.0	13.7	12.4	11.1	38.3	37.0	35.7	34.4	33.1	31.9	30.6	29.3	28.0	26.7	25.4	24.1	
	-0.84	.20	39.9	38.6	37.3	36.0	34.7	33.4	32.1	30.8	29.5	28.2	27.0	25.7	24.4	23.1	21.8	20.5	19.2	17.9	16.6	15.3	14.0	41.2	39.9	38.6	37.3	36.1	34.8	33.5	32.2	30.9	29.6	28.3	27.0	
	0	.50	45.4	44.1	42.8	41.5	40.2	38.9	37.7	36.4	35.1	33.8	32.5	31.2	29.9	28.6	27.3	26.0	24.7	23.4	22.1	20.8	19.5	46.8	45.5	44.2	42.9	41.6	40.3	39.0	37.7	36.4	35.1	33.9	32.6	
	0.84	.80	50.9	49.6	48.4	47.1	45.8	44.5	43.2	41.9	40.6	39.3	38.0	36.7	35.4	34.1	32.8	31.5	30.2	28.9	27.6	26.3	25.0	52.3	51.0	49.7	48.4	47.1	45.8	44.6	43.3	42.0	40.7	39.4	38.1	
	1.28	.90	53.8	52.5	51.3	50.0	48.7	47.4	46.1	44.8	43.5	42.2	40.9	39.6	38.3	37.0	35.7	34.4	33.1	31.8	30.5	29.2	27.9	55.2	53.9	52.6	51.3	50.0	48.7	47.5	46.2	44.9	43.6	42.3	41.0	
	1.64	.95	56.2	54.9	53.6	52.3	51.1	49.8	48.5	47.2	45.9	44.6	43.3	42.0	40.7	39.4	38.1	36.8	35.5	34.2	32.9	31.6	30.3	57.6	56.3	55.0	53.7	52.4	51.1	49.8	48.5	47.3	46.0	44.7	43.4	
	2.33	.99	60.8	59.5	58.2	56.9	55.6	54.3	53.0	51.7	50.4	49.2	47.9	46.6	45.3	44.0	42.7	41.4	40.1	38.8	37.5	36.2	34.9	62.1	60.8	59.5	58.3	57.0	55.7	54.4	53.1	51.8	50.5	49.2	47.9	
	Z	value	25	30	35	40	45	50	55	60	65	70	75	80	80	75	70	65	60	55	50	45	40	35	30	25	30	35	40	45	50	55	60	65	70	75
	-2.33	.01	31.3	30.0	28.7	27.4	26.1	24.8	23.5	22.2	20.9	19.7	18.4	17.1	15.8	14.5	13.2	11.9	10.6	9.3	8.0	6.7	5.4	32.6	31.3	30.0	28.7	27.5	26.2	24.9	23.6	22.3	21.0	19.7	18.4	
	-1.64	.05	35.8	34.5	33.2	31.9	30.7	29.4	28.1	26.8	25.5	24.2	22.9	21.6	20.3	19.0	17.7	16.4	15.1	13.8	12.5	11.2	9.9	37.2	35.9	34.6	33.3	32.0	30.7	29.4	28.1	26.9	25.6	24.3	23.0	
	-1.28	.10	38.2	36.9	35.6	34.3	33.0	31.7	30.4	29.2	27.9	26.6	25.3	24.0	22.7	21.4	20.1	18.8	17.5	16.2	14.9	13.6	12.3	39.5	38.3	37.0	35.7	34.4	33.1	31.8	30.5	29.2	27.9	26.6	25.4	
	-0.84	.20	41.1	39.8	38.5	37.2	35.9	34.6	33.4	32.1	30.8	29.5	28.2	26.9	25.6	24.3	23.0	21.7	20.4	19.1	17.8	16.5	15.2	42.4	41.2	39.9	38.6	37.3	36.0	34.7	33.4	32.1	30.8	29.5	28.3	
	0	.50	46.6	45.3	44.1	42.8	41.5	40.2	38.9	37.6	36.3	35.0	33.7	32.4	31.1	29.8	28.5	27.2	25.9	24.6	23.3	22.0	20.7	48.0	46.7	45.4	44.1	42.8	41.5	40.3	39.0	37.7	36.4	35.1	33.8	
	0.84	.80	52.2	50.9	49.6	48.3	47.0	45.7	44.4	43.1	41.9	40.6	39.3	38.0	36.7	35.4	34.1	32.8	31.5	30.2	28.9	27.6	26.3	53.5	52.2	51.0	49.7	48.4	47.1	45.8	44.5	43.2	41.9	40.6	39.3	
	1.28	.90	55.1	53.8	52.5	51.2	49.9	48.6	47.3	46.0	44.8	43.5	42.2	40.9	39.6	38.3	37.0	35.7	34.4	33.1	31.8	30.5	29.2	56.4	55.1	53.9	52.6	51.3	50.0	48.7	47.4	46.1	44.8	43.5	42.2	
	1.64	.95	57.5	56.2	54.9	53.6	52.3	51.0	49.7	48.4	47.1	45.8	44.6	43.3	42.0	40.7	39.4	38.1	36.8	35.5	34.2	32.9	31.6	58.8	57.5	56.2	54.9	53.7	52.4	51.1	49.8	48.5	47.2	45.9	44.6	
	2.33	.99	62.0	60.7	59.4	58.1	56.8	55.6	54.3	53.0	51.7	50.4	49.1	47.8	46.5	45.2	43.9	42.6	41.3	40.0	38.7	37.4	36.1	63.4	62.1	60.8	59.5	58.2	56.9	55.6	54.3	53.0	51.8	50.5	49.2	

Note. LE = Level of Education.

A possible criticism of the LDST key is that cultural or language differences could influence performance. For example, the combination B-2 (see Appendix Figure 1) is an unfamiliar combination in the Dutch language but might be a known association in English-speaking countries (a bomber plane). Empirical data, however, suggest that the LDST is a culturally robust test (Møller et al., 1998). Houx et al. (2002) administered the LDST in English and Dutch to a large sample of elderly participants ($N=5,804$; age range 70-82 years) in Great Britain, Ireland, and the Netherlands. The authors found a good test-reliability, with a test/re-test correlation of .88 (Pearson's correlation; 2-week test intervals), and only small differences between the countries (Pearson's correlations varying less than .05 for each of the three countries).

The average IQ of our sample was high (see Table 1, average IQ = 113.75; SD = 12.66), which is most likely attributable to the so-called Flynn effect (the gain of IQ over time; Flynn, 1987, 1999) because the GIT norms were collected in the 1980s. Our data also suggest that the association between level of education and IQ is not clear-cut, especially for older participants (see Table 1). Indeed, before the 1960s most people did not go to college or university, but for socioeconomic, rather than intellectual, reasons (Jolles et al., 1995). Nevertheless, we used level of education as a predictor of LDST performance instead of IQ because it is important to use a readily available stratification variable in order to facilitate the use of normative data.

In conclusion, LDST performance is strongly influenced by age, education, and sex. While this study included only cognitively intact, healthy people, in order to generate normative data, the usefulness of the LDST has been demonstrated in clinical samples (e.g. De Groot, Hornstra, Rozendaal, & Jolles, 2003; Dijkstra et al., 2002; Jelicic, Van Boxtel, Houx, & Jolles, 2000; Krabbendam, Van Harten, Picus, & Jolles, 2000; Møller et al., 1998; Ponds, Van Boxtel, & Jolles, 2000; Van Boxtel et al., 1997). The LDST and the normative data provided here can be used in clinical or research settings when information processing speed needs to be measured.

Appendix

W	B	T	P	V	D	G	C	J
1	2	3	4	5	6	7	8	9

T	W	C	G	J	V	B	D	P	V	P	T	D	C	B

5

P	D	V	B	T	D	P	W	B	J	D	T	C	V	G

20

J	P	W	C	B	V	J	D	P	C	G	W	T	B	V

35

T	G	V	B	P	W	C	V	D	J	W	J	G	D	C

50

G	T	J	C	W	C	G	D	J	P	B	V	T	C	B

65

W	P	G	V	B	J	C	P	T	C	G	W	J	D	V

80

J	P	G	D	G	B	J	C	W	V	T	B	D	T	W

95

T	V	G	W	D	P	V	D	B	J	G	T	J	P	B

110

W	C	T	V	P	B	J	G	W	D	V	C	T	P	G

v. 1

Figure 1. The LDST test sheet.

Chapter 8

Assessment of change in performance on the Stroop Color-Word Test, Verbal Learning Test of Rey, and Letter Digit Substitution Test after a test-retest interval of three years

Wim Van der Elst, Martin P. J. Van Boxtel, Gerard J. P. Van Breukelen & Jelle Jolles

Abstract

Precise understanding of test-retest effects and performance instability over time is essential for valid assessment of cognitive change with neuropsychological tests. Unfortunately, such information is lacking for many neuropsychological tests. The present study evaluated the influence of age, gender, and level of education on change in performance after a three year test-retest interval for commonly used test scores of the Verbal Learning Test of Rey (VLT), the Stroop Color-Word Test (SCWT), and the Letter Digit Substitution Test (LDST) in a large sample of people who were aged between 49 and 81 at baseline.

The results showed that test-retest changes differed as a function of age and/or gender for the SCWT Interference, VLT Total recall 1-3, VLT Delayed recall, and LDST measures, but were identical across subgroups for the VLT Immediate recall measure. An Age x Gender interaction was found for the SCWT Interference change score, suggesting that the age-related decline in executive functioning after three years was more pronounced for males than for females. The normal range of change in performance was established, with appropriate corrections for the relevant independent variables. These data provide a useful tool for clinicians in the evaluation of change in the cognitive abilities of individuals.

Introduction

Neuropsychological tests are often administered repeatedly to the same person. For example, a clinician may administer the same test multiple times to monitor change in the cognitive abilities of people who are demented, or a researcher may administer a neuropsychological test repeatedly as part of a randomized clinical trial to evaluate a novel pharmaceutical treatment. When performing serial testing to evaluate deterioration or improvement in cognitive functioning, it is of crucial importance to understand the relation between earlier and later test administrations. Indeed, a serious problem in serial testing is that varying test results may occur even when there was no true change in the cognitive status of a person (Beglinger et al., 2005). This problem is mainly attributable to the less than perfect temporal stability of test scores and the occurrence of test-retest effects such as practice effects (Temkin, Heaton, Grant, & Dikmen, 1999). There are increasing numbers of studies that evaluated the reliability and test-retest changes of neuropsychological tests, but these studies lag considerably behind what is needed in research and clinical practice (Beglinger et al., 2005; Dikmen, Heaton, Grant, & Temkin, 1999; Ivnik et al., 1999). First, reliability estimates and estimates of test-retest effects with the longer test-retest intervals that are used in many clinical and research settings (i.e. months or years rather than days or weeks) are either unavailable or based on small, unrepresentative samples. Secondly, the available studies reported insufficient data: by merely calculating the average changes in test performance over time and test-retest correlations, a clinician or researcher cannot evaluate whether the changes in the test scores of an *individual* are statistically significant or fall within the normal range of expected variation.

The aim of the present study was to establish the normal range of change in performance after 3 years for a large sample of cognitively intact people (age range 49 - 81 at baseline) who were given three often-used neuropsychological tests, i.e. Rey's Verbal Learning Test (VLT; Rey, 1958), the Stroop Color-Word Test (SCWT; Stroop, 1935), and the Letter Digit Substitution Test (LDST; Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2006b). Until now, such data are lacking for these tests. Many clinicians use some idiosyncratic experience-based criterion to determine whether significant change occurred over time (e.g. 'a change of 2 words in the VLT Trial 1 is significant'), but this approach is not optimal and should be replaced by a more methodologically sound approach.

Statistical methods for comparing groups over time are well-developed and straightforward, but there has been serious controversy with respect to the optimal statistical approach to assess change in individuals (see e.g. Collie et al., 2003; Heaton et al., 2001; Temkin et al., 1999). The Reliable Change Index with correction for practice (RCI; Chelune, Naugle, Lüders, Sedlak, & Awad, 1993, see also Jacobson & Truax, 1991) and the multiple regression-based approach (Temkin et al., 1999) are two frequently used methods to assess change, but both methods have some disadvantages. In the RCI method, the predicted

change in performance in an individual equals the mean change score of a normative sample. The RCI method presupposes identical means (and *SDs*) of change for different subgroups as based on age, gender, and other possible relevant variables. However, previous research has shown that the change in performance is not uniform across different subgroups (e.g. Dikmen et al., 1999). In the regression-based approach (see Temkin et al., 1999), the follow-up scores are predicted based on the baseline scores and demographical variables in a normative sample. A problem with the Temkin et al. (1999) approach is that the significance of the change in performance is evaluated by regressing the follow-up scores on the baseline scores (i.e. by using the baseline score as a covariate). Van Breukelen (2006) showed that this approach is biased when there are group differences in the baseline scores (e.g. there are significant differences in the baseline scores of younger versus older people, or lower educated versus higher educated people). The assumption of equal baseline scores is indeed likely to be violated in the Temkin et al. (1999) approach, because it has been consistently shown in previous research that factors such as age, gender, and level of education affect baseline performance on most neuropsychological tests profoundly (Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2005, 2006a, 2006b, 2006c). When follow-up scores are analyzed with the baseline score as covariate and in the presence of group differences in the baseline scores, a phenomenon known as Lord's ANCOVA paradox (Lord, 1967, 1969) is encountered. For example, when the mean difference between the follow-up scores of males and females are equal to the mean difference between the baseline scores of males and females, a significant effect of gender on the follow-up measure will nevertheless be found. Van Breukelen (2006) showed that the approach of using change scores (= follow-up score - baseline score), instead of predicting the follow-up scores with baseline score as covariate, avoids this problem. Thus, in the present study, we evaluated the influence of age, gender, level of education, and their interactions on the change scores in the Verbal Learning Test of Rey (VLT), the Stroop Color-Word Test (SCWT), and the Letter Digit Substitution Test (LDST). Normative change data were provided so that clinicians can evaluate the significance of the change in performance after 3 years in an individual.

Method

Participants

Data were derived from the Maastricht Aging Study (MAAS), a prospective study into the determinants of cognitive aging. MAAS involves a large group of cognitively intact people ($N = 1,856$ at baseline) over the entire adult age range (24 - 81 years at baseline) who underwent extensive cognitive and medical examinations. Baseline measurements were conducted between 1993 and 1996.

The follow-up frequency in MAAS was dependent on the age at first measurement, i.e. 3 years for people aged 49 years and older. A total sample of 1,048, 776, and 1,047 people aged above 49 were administered the SCWT, VLT, and LDST at baseline, respectively. Not all baseline data were used in the analyses. The data of 12 people were excluded because they scored below 24 on the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975). Data were also excluded due to technical problems that occurred during test administration ($n=46$ for the SCWT, $n=8$ for the VLT, and $n=7$ for the LDST). In addition, the SCWT data of 3 participants were excluded from the analyses because these people made more than 20 errors in the colour-word subtask (on a total of 100 items).

Follow-up measurements were conducted about three years later (mean test-retest interval equalled 3.13 years, SD .24 years). Follow-up data were available for a total of 781, 612, and 805 people for the SCWT, VLT, and LDST, respectively. Drop-out percentages equaled 20.8%, 19.0%, and 21.7% for the respective tests. Again, not all available follow-up data were used in the analyses, as the following exclusion criteria were used: technical problems during test administration ($n=8$ for the SCWT, $n=12$ for the VLT, and $n=5$ for the LDST) and physical or cognitive limitations of the testee that hampered test administration ($n=4$ for the SCWT). Thus, we had complete data at baseline and follow-up for 769, 600, and 800 participants for the SCWT, VLT, and LDST, respectively. These data were analyzed in the present study. Note that a visual and an auditory VLT version was used at baseline in MAAS (the auditory version was administered to one fourth of the sample, and the visual VLT to three-fourths of the sample). At follow-up, all participants were given a visual VLT version. Only the data of participants who were given the visual version on both test occasions were used, because mode of presentation was found to affect test-performance profoundly (Van der Elst et al., 2005). Consequently, the total data sample for the VLT is smaller than the data samples for the SCWT and LDST.

The ethnic background of all participants was Caucasian, and all participants were native Dutch speakers. Level of education (LE) was measured by classifying formal schooling in three groups – those with at most primary education (LE low), those with junior vocational training (LE average), and those with senior vocational or academic training (LE high). Basic demographical data for the sample at baseline and at follow-up are provided in Table 1 for three age groups (49-56 year, 59-66 year, and 69+ year). More details regarding the sample frame, subject inclusion procedure, stratification criteria, attrition, and other aspects of the MAAS study have been described in detail elsewhere (Jolles, Houx, Van Boxtel, & Ponds, 1995).

Table 1. Demographic characteristics of the sample of participants who were administered the SCWT, VLT, and LDST at baseline (left) and at follow-up (right).

Age Group	Baseline										Follow-up											
	Level of Education					Male:Female					Level of Education					Male:Female						
	N	M	SD	Low	Average	High	Ratio	N	M	SD	Low	Average	High	Ratio	N	M	SD	Low	Average	High	Ratio	
SCWT	49-56 year	319	52.85	2.73	134	120	65	161:158	269	55.95	2.77	99	112	58	144:125							
	59-66 year	308	63.00	2.58	155	114	39	156:152	252	66.10	2.59	119	101	32	134:118							
	69+ year	360	73.81	3.49	189	114	55	181:179	248	76.42	3.37	128	82	38	121:127							
	Total	987	63.66	9.18	478	348	159	498:489	769	65.88	8.88	346	295	128	399:370							
VLT	49-56 year	245	52.91	2.72	105	90	50	124:121	208	56.01	2.75	79	84	45	111:97							
	59-66 year	234	63.10	2.58	116	90	28	119:115	195	66.25	2.61	96	78	21	105:90							
	69+ year	277	73.75	3.51	147	85	44	135:142	197	76.38	3.32	98	65	34	94:103							
	Total	756	63.69	9.16	368	265	122	378:378	600	66.03	8.86	273	227	100	310:290							
LDST	49-56 year	325	52.87	2.73	138	122	65	164:161	273	55.96	2.77	101	114	58	148:125							
	59-66 year	319	63.04	2.57	163	117	39	162:157	259	66.14	2.58	125	102	32	136:123							
	69+ year	384	73.89	3.50	205	120	57	195:189	268	76.59	3.43	138	89	41	132:136							
	Total	1028	63.88	9.21	506	359	161	521:507	800	66.17	8.98	364	305	131	416:384							

Note. SCWT = Stroop Color-Word Test; VLT = Verbal Learning Test; LDST = Letter Digit Substitution Test. Data on Level of Education were missing for 2 participants who were administered the SCWT at baseline, for 1 participant who was administered the VLT at baseline, and for 2 participants who were administered the LDST at baseline.

Procedure and instruments

All participants were administered the SCWT, VLT, and LDST individually at the neuropsychological laboratory of the Brain & Behaviour Institute (Maastricht, Netherlands). The SCWT (Stroop, 1935; Van der Elst et al., 2006a) measures cognitive flexibility and control (Uttl & Graf, 1997) or executive functioning (Moering, Schinka, Mortimer, & Graves, 2004). Several SCWT versions are available (e.g. Comalli, Wapner, & Werner, 1962; Golden, 1978; Hammes, 1973; Trenerry, Crosson, DeBoe, & Leber, 1989), with variations in the colour and number of the test items, the number of subtests, and the administration procedure. The Hammes version (1973) was used in the present study. This SCWT version consists of three subtasks. The first subtask shows colour words in random order (red, blue, yellow, green) printed in black ink. Subtask two displays solid colour patches in one of these four basic colours. The third subtask contains colour words printed in an incongruous ink colour, e.g. the word yellow printed in red ink. Each subtask involves 100 stimuli that are evenly distributed in a 10 x 10 matrix on a landscape oriented sheet of white paper (A4 format). The participants were instructed to read the words (subtask one), name the colours (subtask two), and name the ink colour of the printed words (subtask three) as quickly and as accurately as possible. The SCWT Interference measure served as the outcome variable. This measure was calculated by subtracting the average time (measured in seconds) needed to complete the first two subtasks from the time needed to complete the third subtask ($\text{SCWT Interference} = \text{time to complete subtask three} - [\{\text{time to complete subtask one} + \text{time to complete subtask two}\} / 2]$; Van der Elst et al., 2006a). The SCWT versions used at baseline and follow-up were identical.

The VLT (Rey, 1958; Van der Elst et al., 2005) is a frequently used measure of verbal memory. In this test, fifteen words were presented one by one on a computer screen in fixed order, with a presentation time of 1 second and an inter-stimulus interval of 1 second. The first trial was followed by four more trials in which the words were presented in identical order. In each trial, the participants were asked to repeat as many words as possible in any order as soon as the presentation stopped. After a delay of 20 minutes, and unexpectedly for the participants, the instruction was given to recall the words learned once more. Finally, a recognition trial was administered. Outcome variables used in the present study were the number of correctly recalled words in trial 1, the total number of correctly recalled words summed over the first three trials and the number of correctly recalled words after the 20-minute delay. These measures will be referred to as the VLT Immediate recall, VLT Total recall 1-3 and the VLT Delayed recall scores, respectively. Alternate VLT versions (that were considered to be parallel; Brand & Jolles, 1985) were used at follow-up and baseline. Note that we used the Total recall 1-3 score instead of the more frequently used Total recall 1-5 score because ceiling effects are often observed in the VLT: the increase in the number of words recalled is especially pronounced

during the first three learning trials, which makes the scores for the last two learning trials to some extent redundant (Van der Elst et al., 2005).

The LDST (Jolles et al., 1995; Van der Elst et al., 2006b) is a measure of general speed of information processing. It is procedurally identical to earlier developed substitution tests such as the Digit Symbol Substitution Test (Wechsler, 1955, 1981), but uses over-learned signs instead of the symbols used in other substitution tests. The LDST key gives the numbers 1 to 9, each of them paired with a different letter. Participants were required to replace the randomized letters with the appropriate digit indicated by the key, as fast and accurately as possible. The number of correct substitutions made in 60 seconds served as the dependent variable. The LDST versions used at baseline and follow-up were identical.

Data analysis

The test scores at baseline and at follow-up will be referred to by adding the subscripts b and f to the variable's names, respectively (i.e. Score_b and Score_f). Test-retest reliabilities of the test scores were estimated by calculating the Pearson correlations between the Score_b and Score_f values (note that the correlations between the VLT_b and VLT_f scores are actually no real test-retest correlations because parallel test versions were used: such correlations are usually referred to as 'coefficients of equivalence'; Franzen, Robbins, & Sawicki, 1989). The change scores were expressed in effect size units (Cohen's d ; Cohen, 1992) to allow comparison over the various measures.

Next, the raw change scores ($= \text{Score}_f - \text{Score}_b$) were regressed on Age, Age^2 , Gender, Level of Education, and all possible two-way interactions (the full models). Age was centered ($\text{Age} = \text{calendar age} - 62.5$) before computing the quadratic terms and interactions to avoid multicollinearity (Marquardt, 1980). Gender was dummy coded with male = 1 and female = 0. Level of Education (LE) was dummy coded with two dummies (LE low and LE high) and LE average as the reference category. The full models were then reduced in a stepwise fashion by eliminating the least significant predictor if its two-tailed p -value was above .005, keeping lower order effects (e.g. age) in the model as long as they were involved in a higher order term (e.g. Age^2 or Age x Gender). For each model, the assumptions of homoscedasticity, normal distribution of the residuals, absence of multicollinearity, and absence of 'influential cases' were checked. Homoscedasticity was evaluated by grouping the participants into quartiles of the predicted scores and applying the Levene test to the residuals. Normal distribution of the standardized residuals was investigated by conducting Kolmogorov-Smirnov tests on the residual values. The occurrence of multicollinearity was checked by calculating the Variance Inflation Factors (VIFs), which should not exceed 10 (Belsley, Kuh, & Welsch, 1980). Cook's distances were calculated to identify possible influential cases (Cook & Weisberg, 1982).

Three steps are conducted to evaluate whether significant change had occurred in an individual. First, the predicted change scores are calculated by using the final regression models, i.e. predicted change score = $B_0 + B_1X_1 + \dots + B_nX_n$. Secondly, the standardized residuals of the change scores are calculated, i.e. $(\text{sign}) [\text{observed change score} - \text{predicted change score}] / SD(\text{residual})$. The (sign) equals +1 in case a higher raw test score indicates better performance (i.e. the LDST and VLT measures), and -1 in case a higher score means worse performance (i.e. the SCWT measure). Thirdly, the standardized residuals of the change scores are compared with the 90% Confidence Interval (CI) of the standardized predicted change scores. If the standardized residual of an individual's score falls between the lower and upper ends of the 90% CI (i.e. between -1.645 and 1.645), no significant change has occurred. If it falls below (or above) the lower end of the 90% CI, significant deterioration (or improvement) had occurred. Note that the value 1.645 comes from the standard normal distribution (i.e. 5% deterioration, 5% improvement). In case the standardized residuals of the change scores in the normative sample are not normally distributed, the values -1.645 and 1.645 should be replaced by the observed 5th and 95th percentiles of the distribution of the standardized residuals, respectively.

The three-step procedure to evaluate whether significant change had occurred in an individual is quite cumbersome and is susceptible to making computational errors. Therefore, a normative table was also established so that the change in an individual's scores can be evaluated without making any calculations. In this table, the lower and upper ends of the 90% CI for the *raw* change score are presented. The raw change scores are calculated as (sign) (Score_f - Score_b), with (sign) equalling +1 in case a higher raw test score indicates better performance (i.e. the LDST and VLT measures), and -1 in case a higher score means worse performance (i.e. the SCWT measure). In case the individual's observed raw change scores falls below (or above) the lower (or upper) end of the CI, significant deterioration (or improvement) had occurred.

Results

Reliability and effect sizes

As shown in Table 2, significant test-retest correlations were observed for all measures (all $ps < .001$). The reliability was high for the LDST score ($r > .85$), acceptable for the SCWT Interference, the VLT Total recall 1-3 and the VLT Delayed recall scores ($r > .60$), and poor for the VLT Immediate recall score ($r < .50$). Effect sizes were generally small, with the exception of the medium effect size of the VLT Immediate recall and the VLT Total recall 1-3 scores (i.e. $d \geq .50$; Cohen, 1992).

Table 2. Test-retest correlations and effect sizes (Cohen's d) of the SCWT, VLT, and LDST measures.

Score	$r(\text{Score}_b, \text{Score}_t)$	Cohen's d
SCWT Interference	.68*	0.18
VLT Immediate recall	.45*	0.57
VLT Total recall 1-3	.62*	0.50
VLT Delayed recall	.66*	0.34
LDST	.87*	0.09

Note. Cohen's $d = (\text{Score}_t - \text{Score}_b) / \text{pooled } SD$.

* $p < .001$ (two tailed)

The multiple linear regression-based approach to assess change

The final linear regression models for the $\sqrt[3]{(\text{SCWT Interference change} + 100)}$, VLT Total recall 1-3, VLT Delayed recall change and LDST change measures are shown in Table 3. The SCWT Interference change score was square root transformed because preliminary data analyses showed that the residuals were positively skewed before transformation of the scores (data not shown). A constant of 100 was added because negative SCWT Interference change scores occurred in the normative sample.

There was no multicollinearity for any of these models (maximum VIF for the final models = 2.009). No severe influence of outliers was found for any of the models (maximum Cook's $d = .276$). The Levene test rejected the homogeneity of variance assumption for the model of the $\sqrt[3]{(\text{SCWT Interference Change} + 100)}$ score. This was taken into account by calculating the $SD(\text{residual})$ s per quartile of the predicted $\sqrt[3]{(\text{SCWT Interference Change} + 100)}$ scores (Van Breukelen & Vlaeyen, 2005), i.e. $SD(\text{residual})$ equals .621 for predicted scores below 9.997, .697 for predicted scores between 9.998 and 10.101, .849 for predicted scores between 10.102 and 10.207, and 1.251 for predicted scores above 10.208. The $SD(\text{residual})$ s to be used in the standardization of the VLT Total recall 1-3, VLT Delayed recall and the LDST change scores equal 4.633, 2.381 and 3.648, respectively. The standardized residuals for the $\sqrt[3]{(\text{SCWT Interference change} + 100)}$ scores were not normally distributed (p of the Kolmogorov-Smirnov $Z < .005$). Thus, the 5th and 95th percentiles of the distribution of the observed standardized residuals in the normative sample should be used to establish the 90% CI of the standardized predicted change scores instead of the ± 1.645 values in the evaluation of the change for the $\sqrt[3]{(\text{SCWT Interference change} + 100)}$ score, i.e. - 1.413 (5th percentile) and 1.401 (95th percentile).

Table 3 presents no regression model for the VLT Immediate recall change score, because none of the predictors reached significance. Thus, the regression model that predicts the change for this score can be written as: predicted VLT Immediate recall change = B_0 , with B_0 equalling the mean change score that was

observed in the normative sample, i.e. 1.08. The *SD*(residual) to be used to standardize the residual of the VLT Immediate recall change score equals the observed *SD* of the VLT Immediate recall change score in the normative sample, i.e. 1.91. Thus, the standardized change score is calculated as is done in the RCI method, that is, $Z = (\text{observed change} - \text{mean change in the normative sample}) / SD(\text{change in the normative sample})$. In other words, the RCI method is a special case of the regression-based approach, i.e. the results of the regression-based and the RCI approaches are identical in case the mean changes (and *SDs*) in scores are equal across subgroups (i.e. none of the predictors in the regression model significantly affect the predicted change score). Note that the standardized residuals of the VLT Immediate recall change score were not normally distributed in the normative sample (p of the Kolmogorov-Smirnov $Z < .005$). Thus, the 5th and 95th percentiles of the distribution of the observed standardized residuals should be used to establish the 90% CI of the standardized predicted change score, i.e. - 1.613 (5th percentile) and 1.529 (95th percentile).

Table 3. Final multiple linear regression models for the $\sqrt{(\text{SCWT Interference} + 100)}$, VLT Total recall 1-3, VLT Delayed recall, and LDST change scores; the full models included Age, Age², Gender, LE low, LE high, and all two-way interactions as predictors.

Change score	Variable	<i>B</i>	Std. Error <i>B</i>	<i>T</i>	Standardized <i>B</i>	<i>R</i> ²	MS(Error)
$\sqrt{(\text{SCWT Interference} + 100)}$	(constant)	10.065	.047	215.557**			
	Age	.008	.005	1.628	.081		
	Gender	.166	.065	2.566	.090		
	Age x Gender	.028	.007	3.809**	.189	.068	.791
VLT Total recall 1-3	(constant)	2.854	.189	15.061**			
	Age	-.098	.021	-4.591**	-.185	.032	21.505
VLT Delayed recall	(constant)	1.012	.098	10.373**			
	Age	-.036	.011	-3.248**	-.132	.016	5.680
LDST	(constant)	.953	.189	5.038**			
	Age	-.069	.015	-4.672**	-.166		
	Age ²	-.005	.002	-3.021*	-.107	.044	13.341

Note. Age was centered by subtracting 62.5 from the calendar age at baseline; Gender was coded as Male = 1, Female = 0.

* $p < .005$; ** $p < .001$

The final model for the $\sqrt{(\text{SCWT Interference change} + 100)}$ score showed that the age-related change in performance depended on gender (see Table 3 and Figure 1). Note that Figure 1 presents the raw SCWT Interference change score instead of the $\sqrt{(\text{SCWT Interference} + 100)}$ scores, for reasons of clarity. Persons aged 55 or higher at baseline showed an increase in the SCWT Interference change score (i.e. worse performance at follow-up), and this change was much larger for males than for females. The predicted VLT Total recall 1-3 and VLT Delayed recall change scores were significantly affected by age (see Table 3 and Figures 2 and 3): the predicted VLT Total recall 1-3 and VLT Delayed recall change scores decreased from about 4 and 1.5 words for persons aged 50

years at baseline to 1 and 0.4 words for persons aged 80 at baseline. The predicted LDST change score was affected by a linear as well as a quadratic age-effect. As shown in Figure 4, the predicted LDST change scores remained more or less constant for people aged between 50 and 60 years, and then declined strongly though the change remained positive (i.e. better performance at follow-up as compared to baseline performance) until the age of about 70 years (at baseline). Level of education did not affect any of the change scores.

Figure 1. Predicted SCWT Interference change scores as a function of age at baseline, for males and females.

Figure 2. Predicted VLT Total recall 1-3 change scores as a function of age at baseline.

Figure 3. Predicted VLT Delayed recall change scores as a function of age at baseline.

Figure 4. Predicted LDST change scores as a function of age at baseline.

As described earlier, the regression-based approach to assess change in an individual consists of three steps. For example, let us consider a 50-year-old person who obtained a VLT Delayed recall_b score of 10 and a VLT Delayed recall_f score of 7. Thus, the VLT Delayed recall change score equals -3 (= 7 - 10). Is a decline of 3 words in the VLT Delayed recall score significant for a 50-year-old person? First, the predicted VLT Delayed recall change score is calculated, which equals 1.462 (= $1.012 + [-.036 * \{50 - 62.5\}]$). Secondly, the standardized residual of the change score is calculated, which equals -1.874 (= $[-3 - 1.462] / 2.381$). Thirdly, it is concluded that significant deterioration had occurred because the standardized residual of the change score (-1.874) falls below the lower end of the 90% CI of the standardized predicted change score (i.e. below -1.645).

A user-friendly normative table

Clinicians may hesitate to use the three-step normative procedure to evaluate whether significant change in an individual's scores had occurred because this procedure requires quite some calculation, and there is room for error in conducting these computations. Therefore, we used the regression models in Table 3 to compute a normative table that presents the 90% CIs for the *raw* SCWT Interference change, VLT Immediate recall, VLT Total recall 1-3, VLT Delayed recall, and LDST change scores, stratified by their relevant predictors (Table 4). Note that use of the normative table instead of using the three-step procedure described above is at the expense of some accuracy because interpolation is required when an individual is not exactly 50, 52.5, 55, ..., 80 years old at baseline (due to space constraints, it is not possible to tabulate data for each value of the predictor age).

Use of Table 4 is straightforward. For example, let us consider a 60-year-old male who obtained a score of 45, 8, 24, 5, and 35 for the SCWT Interference_b, VLT Immediate recall_b, VLT Total recall 1-3_b, VLT Delayed recall_b and LDST_b measures. Follow-up scores equalled 75, 7, 22, 3, and 25 for the respective

scores. Thus, the SCWT Interference change, VLT Immediate recall change, VLT Total recall 1-3 change, VLT Delayed recall change and LDST change scores equalled -30, -1, -2, -2, and -10, respectively. Significant deterioration had occurred for the SCWT Interference score because the observed change (i.e. -30) was below the lower end of the 90% CI presented in Table 4 (i.e. -28.6). The changes for the VLT Immediate recall, VLT Total recall 1-3 and the VLT Delayed recall scores were not significant because they were within the lower and upper ends of the 90% CI. Significant deterioration also occurred for the LDST score because the observed change (i.e. -10) was below the lower end of the 90% CI (i.e. -4.9).

Discussion

The present study evaluated the change in performance after 3 years in the SCWT, VLT and LDST in a large sample of cognitively intact people, aged between 49 and 81 years at baseline. The results showed that the test-retest changes were identical for different subgroups as based on demographical variables for the VLT Immediate recall score, but differed as a function of age and/or gender for the SCWT Interference, VLT Total recall 1-3, VLT Delayed recall, and LDST change scores. For the VLT Immediate recall change score, a positive and constant change of 1.08 was predicted for people across the entire age range, which means that about one word more was recalled at follow-up as compared to baseline. The VLT Total recall 1-3 and VLT Delayed recall change scores were also predicted to be positive for people across the entire age range studied, but the predicted change was smaller for older people as compared to the predicted change for younger people (see Figures 2 and 3). A positive LDST change score was predicted for people who were aged between 50 and 70 at baseline, but the predicted change for people who were older than 70 at baseline was negative. Though it has been frequently advocated that measures of information processing speed such as the LDST are amongst the first to decline with increasing age (see e.g. Salthouse, 1996), the results of the present study do not support this claim. Our results rather support the more recent theories that highlight the fundamental role of executive functioning in cognitive aging (see West, 1996). Indeed, the age-related change in the SCWT Interference score (a measure of executive functioning) was quite pronounced for people who were aged above 55, though the age-related decline in performance was larger for males than for females (see Figure 1). The interaction between age and gender on the SCWT Interference change score was not expected because most longitudinal studies report either no differences in the age-related decline in cognitive functions for men and women (Aartsen, Martin, & Zimprich, 2004), or, in case gender differences are observed, a more pronounced decline for females (Meinz & Salthouse, 1998). On the other hand, volumetric imaging studies showed a stronger effect of aging on the frontal lobes of the brain of men in comparison

Table 4. Normative table for the raw SCWT Interference, VLT Immediate recall, VLT Total recall 1-3, VLT Delayed recall, and LDST change scores. The lower and upper ends of the 90% Confidence Intervals for the raw change scores are presented, stratified by their relevant predictors.

		Age (at baseline)													
		90% CI	50	52.5	55	57.5	60	62.5	65	67.5	70	72.5	75	77.5	80
SCWT Interference	Males	Lower end	-13.6	-15.5	-17.5	-21.8	-28.6	-44.0	-46.1	-48.3	-50.5	-52.7	-55.0	-57.2	-59.5
		Upper end	20.6	19.0	17.4	17.7	19.9	28.1	26.6	25.0	23.5	21.9	20.3	18.7	17.0
	Females	Lower end	-17.6	-18.0	-20.8	-21.2	-21.7	-22.1	-22.5	-27.8	-28.2	-28.7	-29.2	-29.6	-30.1
		Upper end	17.3	16.9	18.5	18.1	17.8	17.4	17.0	20.5	20.2	19.8	19.4	19.1	18.7
VLT Immediate recall	Males & Females	90% CI	All ages												
		Lower end	-2.0												
		Upper end	4.0												
VLT Total recall 1-3	Males & Females	90% CI													
		Lower end	-3.54	-3.79	-4.03	-4.28	-4.52	-4.77	-5.01	-5.26	-5.50	-5.75	-5.99	-6.24	-6.48
		Upper end	11.70	11.46	11.21	10.97	10.72	10.48	10.23	9.99	9.74	9.50	9.25	9.01	8.76
VLT Delayed recall	Males & Females	90% CI	50	52.5	55	57.5	60	62.5	65	67.5	70	72.5	75	77.5	80
		Lower end	-2.5	-2.5	-2.6	-2.7	-2.8	-2.9	-3.0	-3.1	-3.2	-3.3	-3.4	-3.4	-3.5
		Upper end	5.4	5.3	5.2	5.1	5.0	4.9	4.8	4.7	4.7	4.6	4.5	4.4	4.3
LDST	Males & Females	90% CI	50	52.5	55	57.5	60	62.5	65	67.5	70	72.5	75	77.5	80
		Lower end	-5.0	-4.9	-4.8	-4.8	-4.9	-5.0	-5.3	-5.5	-5.8	-6.2	-6.7	-7.2	-7.8
		Upper end	7.0	7.1	7.2	7.2	7.1	7.0	6.8	6.5	6.2	5.8	5.3	4.8	4.2

Note. The raw change scores are calculated as (sign) (Score_t - Score_b), with (sign) equaling + 1 in case a higher raw test score indicates better performance (i.e. the LDST and VLT measures), and - 1 in case a higher score means worse performance (i.e. the SCWT measure). Significant deterioration (or improvement) had occurred in case the individual's observed raw change score falls below (or above) the lower (or upper) end of the 90% CI.

to females (Cowell et al., 1994; Murphey et al., 1996), which may explain the Age x Gender interaction for the SCWT Interference change score.

We used a regression-based approach to assess the significance of change in performance in individuals. This approach has the advantage that it allows for a correction of demographical influences in the prediction of the change scores. As noted, such a correction was necessary for the SCWT Interference, VLT Total recall 1-3, VLT Delayed recall, and LDST change scores, but not for the VLT Immediate recall change score. The RCI approach can be seen as a special case of the regression-based approach that is only valid in case the means and *SDs* of change for different subgroups are identical. We recommend use of the regression-based approach for future studies on the assessment of change in individuals, because it is more generally applicable than the RCI approach.

This brings us to some general remarks and limitations of the present study. First, educational level did not affect any of the change scores considered in the present study. Many cross-sectional studies have shown that level of education is correlated with individual differences in cognitive functions (e.g. Van der Elst et al., 2005, 2006a, 2006b, 2006c), but correlations between educational level and (cross-sectional) test scores are not sufficient to conclude that education is a predictor of cognitive *change*. Indeed, longitudinal studies such as the present one are needed to evaluate predictors of cognitive change. Anstey and Christensen (2000) reviewed longitudinal studies that evaluated the effect of educational level on cognitive change. They concluded that the rate of decline of higher educated people was less rapid for tests that measure crystallized abilities (i.e. cognitive abilities that depend upon the accumulation of formal and informal educational experiences; LaRue, 1992), memory, and general mental status (as measured with the MMSE), whilst educational level was less significant for predicting change in fluid abilities (i.e. abilities reflect maturational growth and decline of neuronal structures; LaRue, 1992) and speed. Thus, an effect of education on the VLT Immediate recall, VLT Total recall 1-3 and the VLT Delayed recall change scores was to be expected. The lack of educational effects on the change of memory measures that were considered in the present study may be related to the fact that quite stringent inclusion criteria were used (i.e. only cognitively intact people were included in the sample). It is also conceivable that the effects of education on change of memory measures are quite small and may only be detected when a test-retest interval of more than three years is used.

Secondly, the results of the present study suggested adequate to good reliability for the SCWT Interference, VLT Total recall 1-3, VLT Delayed recall, and LDST scores (all $r_s > .60$), but the reliability of the VLT Immediate recall score was poor ($r = .45$). It should be noted that the approach of calculating test-retest correlations (or coefficients of equivalence) as estimates of reliability presupposes that stable constructs are measured (i.e. constructs that do not change from baseline to follow-up). This assumption may not have been fulfilled in the present study, because a fixed test-retest interval of three years was used. Consequently, the reliability estimates that were reported in the present

study were slightly lower than the ones that are generally reported in studies that used shorter test-retest intervals (see e.g. Beglinger et al., 2005; Houx et al., 2002; Rasmussen et al., 2004; Schmidt, 1996). Note however that the reliability of the VLT Immediate recall score is consistently found to be low, also in studies with short test-retest intervals (Schmidt, 1996). The poor reliability of the VLT Immediate recall score may partly explain why none of the demographic variables were significantly predictive the VLT Immediate recall change score. After all, measurement error variance in a change score is the sum of the measurement error variances at baseline and at follow-up, and so the change in VLT Immediate recall had a low reliability.

Thirdly, the test performance was better at follow-up as compared to baseline for people aged between 50 and 55 for all the test scores considered in the present study (and for most tests also for older people, see above). The better performance at follow-up as compared to baseline suggests that practice effects did occur, at least in the younger subgroup of our sample. It has been consistently shown in other studies that improvement occurs from baseline to follow-up, especially in memory tests (Schmidt, 1996). For this reason, alternate VLT versions were used at baseline and at follow-up. Although using parallel tests avoids the obvious problem of remembering the content of a test, procedural learning cannot be circumvented (Benedict & Zgaljardic, 1998; Claus, Mohr, & Chase, 1991; Schmidt, 1996). As a consequence, better performance at follow-up as compared to baseline performance still may be found. For the SCWT and LDST, no parallel test versions were used (because it was expected to be unlikely that participants would remember the content of these tests), but the fact that at least a subgroup of our sample scored higher at follow-up as compared to baseline suggests that some procedural learning had also occurred in these tests (see Figures 1 and 4).

Fourthly, due to the long test-retest interval that was used in the present study and the occurrence of practice effects (i.e. procedural learning), it is difficult to attribute the age-related changes in performance to ‘true changes’ in the underlying constructs that are measured by the neuropsychological tests used in the present study. Indeed, the observed change scores most likely reflect the ‘net effect’ of the positive effects of procedural learning on test-performance at follow-up and the negative effects of a true age-related decline in the underlying cognitive abilities. It is not possible to distinguish the effects of ‘true change’ from the effects of procedural learning in the present study, which is a common problem in most longitudinal studies that use repeated measurements of cognitive tests. This is however not relevant for the practical use of the present norms because a procedural learning effect may be expected to occur in practice just as well as in our reference sample.

Fifthly, the drop-out percentage was about 20% in the present study. Van Beijsterveldt and colleagues (2002) found that the drop-out in MAAS was related to certain characteristics of the participants (e.g. lower educated people dropped-out more often than higher educated people). Non-random attrition

may lead to biased parameter estimates of cognitive change, but Van Beijsterveldt and colleagues (2002) found little effect of attrition on the estimates of change. Their results suggest that drop-out did not significantly affect the parameter estimates reported in the present study (see Table 3), but caution is required because evaluating whether attrition affects the estimates of change is a cumbersome task that requires several assumptions that may not have been fulfilled (see Van Beijsterveldt et al., 2002).

Sixthly, a test-retest interval of 3 years was used in the present study. Other studies reported that test-retest changes and reliability estimates vary as a function of the length of the test interval (Dikmen et al., 1991; McCaffrey, Ortega, & Haase, 1993). Thus, the norms that were established in the present study should only be used to evaluate change in performance when a test-retest interval of about 3 years was used. Further research is needed to determine whether these norms are also suitable to evaluate the change in people who were tested with a shorter or longer test interval. This requires a large sample of people who were retested with variable intervals: by adding the length of the test-retest interval as an extra predictor in the regression-models, the influence of the length of the interval on change can be determined and corrected for if necessary.

Chapter 9

Concluding remarks

The first objective of this thesis was to evaluate the influence of age and age-extrinsic factors on performance of cognitive tests. By using logistic and linear regression analysis, the relevance and relative contribution of demographical factors on inter-individual differences in cognitive abilities and the change in cognitive abilities over time was formally modelled. The relationships between demographical variables and cognitive abilities that are observed in normative studies are robust due to the fact that large, representative samples are used (Salthouse, 1996). Consequently, such formal models increase the insight in theoretical issues regarding the onset, patterns, and magnitude of age-related changes and differences in cognition.

The second objective was to establish the normal range of performance for three commonly used (i.e. the Stroop Color-Word Test, the Verbal Learning Test, and the Verbal Fluency Test) and three new (i.e. the Concept Shifting Test, the Paper & Pencil Memory Scanning Test, and the Letter digit substitution Test) neuropsychological tests. These neuropsychological tests were selected rather than other tests because they quantify cognitive abilities that are typically assessed in most neuropsychological examinations.

This last chapter provides a general discussion of some of the results that were reported in this thesis, on methodological issues, and on relevant aspects of neuropsychological assessment that were not studied in the present thesis. In addition, some recommendations for improving the practice of neuropsychological assessment in the future are given.

Cognitive reserve and aging

The concept of ‘cognitive reserve’ (Satz, 1993) is widely used to explain why the effects of neurodegenerative changes that are associated with brain pathology or aging vary considerably across individuals (e.g. Cummings et al., 1998; Stern et al., 1996; Scarmeas & Stern, 2003, 2004). Cognitive reserve is the hypothesized capacity of a person to sustain the effects of brain damage: people with more reserve capacity are assumed to be less vulnerable to the damaging effects of brain injuries or to the changes in the brain that are associ-

ated with aging, as compared to people with less reserve capacity. Thus, the cognitive reserve concept assumes that there are inter-individual differences in reserve capacity, and that there is some critical threshold in reserve capacity that needs to be exceeded before clinical or functional deficits emerge. Stern (2003) distinguished active and passive models of reserve. Passive components of cognitive reserve involve brain structures that add capacity to efficient information processing, retrieval of memories and problem solving. The brains of individuals with more passive reserve are thus assumed to differ from the brains of people with less reserve (e.g. larger number of synapses or larger brain size). Active components of reserve involve experiences such as education, occupational activities, or leisure activities. Due to these experiences, people are assumed to process tasks in a more efficient way and cope better with brain pathology.

In the present thesis, the effect of level of education on performance in cognitive tests was evaluated. Such analyses are not only useful in the context of normative research but can also be considered as tests of the cognitive reserve hypothesis because education is often regarded as a proxy of cognitive reserve and low performance on cognitive tests is often considered as an indication of brain damage. The results of the present thesis showed significant effects of level of education for the large majority of test scores that were considered in the cross-sectional analyses (Chapters 2 – 7). The test performance was especially influenced by the contrast between low and average level of education, as opposed to average and high education. For example, the effect of a low level of education as compared with an average level of education on a test that measures information processing speed (LDST) was comparable to about 15 years of aging, whilst the effect of an average as compared with a high educational level was about 5 years. Thus, the probability of exceeding some critical threshold (i.e. low test scores) was larger in lower educated people as compared to their higher educated counterparts, which seems to be in line with the predictions from the cognitive reserve theory. On the other hand, many authors have suggested that the test-performance of people with a lower education should drop faster as compared to their higher educated counterparts to provide support for the cognitive reserve hypothesis (e.g. Schmand et al., 1997a), that is, an interaction between age and level of education on test performance should be observed. Only three Age x Level of Education interactions on test performance were found in the present thesis. The interactions for the SCWT Interference score and the SCWT error score for the colour-word trial were in the expected direction (i.e. a more pronounced age-related decline in lower educated people as compared to their higher educated counterparts), but the interaction for the VLT Trial 1 score was in the opposite direction. It is unclear to what factors this latter interaction should be attributed to.

Cross-sectional designs have the disadvantage that age-effects are confounded by cohort effects. Thus, the effects of level of education and the interactions between age and level of education that were observed with the cross-sectional studies (Chapters 2 – 7) may be considered as circumstantial evidence

for the cognitive reserve hypothesis, but these findings should be confirmed by longitudinal studies. However, no evidence for effects of education on cognitive change was found in the longitudinal analyses in the present thesis (Chapter 8). This lack of effects of level of education on the age-related change in the performance of cognitive tests after 3 years may be related to several factors. First, only cognitively intact people were included in the MAAS sample at baseline. It is possible that an effect of level of education on cognitive change would have been found in case a more heterogeneous sample was used (e.g. a sample that also includes people with cognitive deficits). Secondly, the effects of level of education on change of cognitive measures may be relatively small and may only be detected when a test-retest interval of more than three years is used. Thirdly, educational level was used as a proxy of brain reserve, but other proxy measures such as intelligence may reflect cognitive reserve more directly than level of education does (Satz, 1993). For example, Schmand and colleagues (1997b) showed that low IQ predicted the incidence of dementia better than low level of education.

In conclusion, it may be overly simplistic to consider the results of the present study as being favourable for the cognitive reserve hypothesis. The longitudinal analyses showed no effects of education on cognitive change, and although strong main effects of level of education were found in the cross-sectional analyses, it is not entirely clear whether a main effect of education can be considered as evidence for the cognitive reserve hypothesis. The main problem with the cognitive reserve theory is that although it is a useful conceptual framework or metaphor, the concept of cognitive reserve itself as well as the predictions that arise from this theory are only loosely defined (i.e. are main effects of education sufficient or is an Age x Education interaction required?). Until now, most research on cognitive reserve used variables such as education or occupational activity as proxies of cognitive reserve. In doing so, other possible relevant predictors of age-related cognitive change may be overlooked. For example, Anstey and Christensen (2000) noted that only a small number of studies evaluated the influence of health status on cognition. The studies that did evaluate the effects of health on cognition, suggested that health affected cognitive decline (e.g. Christensen et al., 1994; Whitfield et al., 1997). More research is needed to evaluate the effect of health on cognition, because it is possible that the effects of education or level of occupational activity are moderated by – or interact with – factors such as health status (Meijer et al., 2006).

Regression-based versus traditional normative procedures

In the present thesis, we used a regression-based approach to establish the normal range of performance on neuropsychological tests. As compared to the traditional approach of establishing norms, the regression-based procedure has some significant methodological advantages, as was detailed in Chapter 2.

Some additional advantages of the regression-based method in the construction of norms can be given.

First, regression-based norming can handle continuous, categorical and/or dummy variables as independent variables. The traditional normative approach cannot handle continuous variables such as age. These variables need to be categorized, which reduces measurement information and statistical power (MacCallum et al., 2002). Categorization of an independent variable is based on the implicit assumption that people within each subgroup do not differ significantly from each other with respect to the dependent variable of interest. This assumption is often problematic. For example, suppose that age was categorized as 21 – 30, 31 – 40, ..., 71 – 80 years. It is possible that people who are in the age group 21 – 30 do not differ much with respect to the cognitive variables of interest, but it is unlikely that the same applies to people who are in the age group 71 – 80 years.

Secondly, the regression-based approach allows for correcting the norms for a larger number of independent variables. In the traditional normative approach, a minimum of about 90 participants is required per subgroup (Gibbons Natrella, 1963), which places heavy restrictions on the possibilities to correct norms for multiple independent variables. For example, it is almost impossible from a practical viewpoint to provide traditional norms that are corrected for age group (12 categories), gender (2 categories), level of education (3 categories) and handedness (2 categories), as this would require a normative sample of 12,960 participants ($= 12 \times 2 \times 3 \times 2 \times 90$). For each independent variable that is added, a huge increase in the required number of participants occurs (i.e. 12,960 multiplied by the number of categories of the additional independent variable). In the regression-based approach, such limitations are much less pronounced.

Thirdly, the regression-based approach provides a general framework for establishing norms that applies to many different situations: norms can be provided for (semi)continuous outcome variables, for categorical outcome variables after dichotomization, for dichotomous outcome variables, and for change in outcome variables over time. In case of complex datasets with multiple measurement occasions and missing values, the basic method can be extended to a multilevel or mixed regression model (Van Breukelen & Vlaeyen, 2005).

Viable independent variables in normative research

In the present study, the effects of age, gender, and level of education on cognitive performance were modelled (see the regression-models in Chapters 2 – 8). The goodness-of-fit of these models with the empirical data may be significantly improved by adding other variables to the model (that are roughly uncorrelated with age, gender, and education). Indeed, a large number of independent variables have been suggested to be correlated with performance on cognitive tests, examples of which are health (Christensen et al., 1994; Whit-

field et al., 1997), activity levels (Christensen et al., 1996), sensory functioning (Van Boxtel et al., 2000), aerobic capacity (Van Boxtel et al., 1997), diastolic blood pressure (Cacciatore et al., 1997), and certain apolipoprotein genotypes (Haan et al., 1999). Adding such variables to the regression models may have been interesting from a theoretical viewpoint, but an important limitation with respect to the choice of independent variables in normative studies is that all independent variables should be measured easily and straightforward. Indeed, each variable that is used in the normative analyses also needs to be measured by the end user of the norms (the clinician or the researcher). Therefore, it is not viable to include independent variables that are costly, complex, or time consuming to measure in normative analyses.

This is also the reason why we used level of education as a proxy of general ability in the normative analyses (See Chapters 2 – 8), rather than a variable such as Level of Occupational Activity (LOA). LOA may be a more valid proxy of general ability than level of education, because many people are poorly educated due to socio-economic rather than academic reasons (especially people aged 50+). These people may have ‘proven’ themselves at the workplace and followed courses so that they achieved a professional position that is nowadays only attained by people with higher degrees (Jolles et al., 1995). Despite the fact that LOA may thus be a better proxy of general ability as compared to level of education in certain cases, we nevertheless used level of education because the measurement of LOA is more difficult as it requires an estimation of the complexity, knowledge and required experience of a person’s occupation.

Ecological validity and neuropsychological assessment

In neuropsychological assessment, the performance of people is observed in a standardized test-setting to make inferences about unobserved, or latent, cognitive abilities (Lezak et al., 2004). The question arises to what extent the performance in standardized test settings is predictive with respect to real-life situations. One may wonder, for example, how the score on a verbal memory test such as the VLT relates to real-life tasks that involve memory components, such as the ability to live independently or the ability to work? It is important to address such questions because neuropsychologists often make recommendations based on the performance on neuropsychological tasks that may have far reaching consequences for a person’s life. Especially in a medico-legal context, the relationship between neuropsychological test scores and present and future functioning in real-world settings is often challenged by attorneys in the courtroom (Sbordone, 1996).

The functional and predictive relationship between a person’s performance on (neuro)psychological tests and the person’s behaviour in a variety of real-world settings is defined as the ecological validity of a test (Sbordone, 1996; pp. 16). Two approaches are used to address the problem of ecological validity of

neuropsychological tests: the verisimilitude and the veridicality approach (Chaytor & Schmitter-Edgecombe, 2003). Verisimilitude refers to the degree of which the cognitive demands of a test *theoretically* resemble the cognitive demands in the everyday environment, whilst veridicality is the degree to which existing tests are *empirically* related to everyday functioning (Franzen & Wilhelm, 1996). The typical verisimilitude approach consists of abandoning the existing neuropsychological tests and creating new ones with optimal ecological validity explicitly held in mind when the test is developed. The Behavioural Assessment of Dysexecutive Syndrome (Wilson et al., 1996) is an example of a test that was developed by using the verisimilitude approach to maximize ecological validity, because it consists of subtasks that simulate daily activities such as planning a route on a map. The typical veridicality approach on the other hand does not abandon existing tests but evaluates the ecological validity of existing tests by relating test outcome to various measures of real-world functioning such as employment status or measures of activities of daily life (e.g. Burgess et al., 1998; Farmer & Eakman, 1995; Kibby et al., 1998; LeBlanc et al., 2000).

The problem of the ecological validity of neuropsychological tests was not addressed in the present thesis (nor other aspects of validity), but a project that uses the veridicality approach is at hand⁴. This project aims at developing a 'Cognitive Index' (CI), i.e. a measure for general neuro-cognitive functioning. The CI score will be derived from the tests that are described throughout this thesis (i.e. the VLT, LDST, SCWT, VFT, MST and CST). In the first phase of the project, different operationalizations of the CI will be constructed, based on theoretical and empirical considerations. For example, the processing speed theory (e.g. Salthouse, 1996) claims that information processing speed is fundamental in the age-related decline of cognitive functions. Thus, based on this theory, the CI should depend heavily on measures of information processing speed such as the LDST and subtasks of the SCWT. On the other hand, other theories emphasize the role of executive functions in cognitive aging (e.g. West, 1996). Thus, based on the executive function theory of cognitive aging, the CI should be mainly based on executive function measures such as the SCWT Interference and the VFT measures. Another approach to construct a CI is to determine the relations between various test scores empirically and reduce them to several factors of interest, using statistical techniques such as factor analysis. Alternatively, principles from Item Response Theory can be applied to construct the CI (see below).

Thus, several operationalizations of the CI are conceivable. To determine which operationalization is optimally ecologically valid, the relation between the different CIs and functional autonomy (as measured with the SF-36; Ware & Sherbourne, 1992) will be determined in the second phase of the CI project.

⁴ "Successful Cognitive Aging and the Cognitive Index (CI): A normative study in the Maastricht Aging Study (MAAS)", Van der Elst, W., Van Boxtel, M., & Jolles, J. (NWO/ZonMw grant 014-91-077).

MAAS provides an excellent design to evaluate ecological validity because a large number of people were administered cognitive tests as well as measures of functional autonomy at baseline and after 3, 6, and 9 years. The CI operationalization that is the most predictive for functional autonomy at the different measurement occasions will be restrained. Ideally, the end product of this project will provide neuropsychologists with an ecologically valid measure of general neuro-cognitive functioning, which may be expected to be a useful tool in clinical and research settings.

Neuropsychological assessment: where to go from here?

Several improvements in the methods and the approaches of neuropsychological assessment may arise in the future.

First, use of Item Response Theory (IRT) methods instead of Classical Test Theory (CTT) methods may have several advantages in the assessment of cognitive functions. For example, a serious problem with the CTT is that it assumes linearity of measurement. This means that equal changes in one part of the continuum of the latent construct that is measured by a test (i.e. a cognitive ability) are assumed to result in the same change in the observed score as would an equal change in a different part of the continuum of the construct. Many (neuro)psychological tests have nonlinear measurement (Mungas & Reed, 2000), for example due to floor and ceiling effects. In the context of evaluating change, nonlinear measurement presents problems with measuring the range of change in the underlying construct (Mungas et al., 2003). Use of IRT methods allow for constructing psychometrically matched scales (i.e. scales that have the same reliability over the entire continuum of the latent construct; Chapman & Chapman, 1973) with linear measurement over the range of the continuum of the cognitive abilities that are measured. Such scales may be a significant improvement for the practice of neuropsychological assessment.

Secondly, neuropsychological tests are often administered repeatedly to the same person with variable test intervals in clinical settings, for example to monitor decline in the cognitive abilities of people who are demented or to evaluate the effects of an intervention aimed at improving cognitive performance. In the present thesis, we provided norms for assessing change when a person was tested with a test-retest interval of three years (see Chapter 8). Such norms provide a first step in the assessment of change in the clinic, but norms should also be provided to evaluate change in performance when variable test-retest intervals are used. Such norms could not be provided in the present thesis because the MAAS study uses fixed test-retest intervals, but the regression-based method can be easily adapted for providing norms for variable test-retest intervals. Indeed, when a sample of people has been tested on two occasions with a variable test-retest interval, norms can be established by using the

method described in Chapter 8 with an extra predictor in the multiple linear regression models, i.e. length of the test-retest interval.

Thirdly, regardless how well standardized, reliable and valid a neuropsychological test is, there is never a one-to-one correspondence between a specific test score and a latent cognitive ability to be measured (Dodrill, 1997; Thompson et al., 2005). As a consequence, poor performance on a test may be caused by a deficit in the cognitive ability that the test is assumed to measure, but may also be caused by other factors. Despite the fact that it is acknowledged that all neuropsychological tests are multifactorial, two approaches can be followed to allow for a more straightforward and specific interpretation of the test scores in terms of underlying cognitive abilities. First, existing neuropsychological tests may be modified in such way that some alternative explanations of poor test performance can be excluded (i.e. poor test performance due to other factors than a deficit in the underlying cognitive ability that the test is assumed to measure). An example of this approach is given in Chapter 5, in which the CST was presented. This test was based on the TMT, but several of the problems with the test design of the TMT that hamper interpretation of the test scores were avoided in the CST. Secondly, test paradigms with a solid theoretical background that stem from a cognitive or experimental tradition may be adapted for use in the clinic to distinguish cognitive abilities more precisely. An example of this approach was given in Chapter 6, in which Sternberg's Memory Scanning Test was adapted for use in the clinic. In contrast to other neuropsychological tests that measure working memory, the P&P MST allows for distinguishing non-memory processing stages and memory scanning stages in working memory. Such tests thus allow for a more precise interpretation of the test scores.

In conclusion, the present thesis increased our insight in the factors that affect inter-individual differences in cognitive functioning and cognitive change over time. The relevance and relative contribution of factors such as age, gender and level of education was established for various cognitive tests. The influence of these variables on cognitive functioning was shown to differ across cognitive domains. We provided normative data for neuropsychological tests that measure cognitive abilities that are often tested in clinical and research settings. Until now, appropriate norms for these tests were not available. The norms presented in this thesis provide a valuable tool for many neuropsychologists and researchers in the Netherlands and abroad.

References

- Anstey, K., Stankov, L., & Lord, S. (1993). Primary aging, secondary aging and intelligence. *Psychology and Aging, 8*, 562-570.
- Aartsen M. J., Martin M., & Zimprich D. (2004). Gender differences in level and change in cognitive functioning - Results from the longitudinal aging study Amsterdam. *Gerontology, 50*, 35-38.
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: testing and interpreting interactions*. Newbury Park (California): Sage.
- Anstey, K., & Christensen, H. (2000). Education, activity, health, blood pressure and apolipoprotein E as predictors of cognitive change in old age: a review. *Gerontology, 46*, 163-177.
- Archibald, C. J., & Fisk, J. D. (2000). Information processing efficiency in patients with multiple sclerosis. *Journal of Clinical and Experimental Neuropsychology, 22*, 686-701.
- Ardila, A. (1995). Directions of research in cross-cultural neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 17*, 143-150.
- Auriacombe, S., Fabrigoule, C., Lafont, S., Amieva, H., Jacqmin-Gadda, H., & Dartigues, J. (2001). Letter and category fluency in normal elderly participants: a population based study. *Aging Neuropsychology and Cognition, 2*, 98-108.
- Beets, B. (2000). *Rimpels in Europa*. [on-line]. Available: <http://www.nidi.knaw.nl/web/html/public/demos/dm00102.html>.
- Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., et al. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology, 20*, 517-529.
- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. New York: Wiley.
- Benedict, R., & Zgaljardic, D. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical & Experimental Neuropsychology, 20*, 339-352.
- Benito-Cuadrado, M. M., Esteba-Castillo, S., Böhm, P., Cejundo-Bolivar, J., & Pena-Casanova, J. (2002). Semantic verbal fluency in animals: A normative and predictive study in a Spanish population. *Journal of the Clinical and Experimental Neuropsychology, 24*, 1117-1122.
- Birren, J. E. & Schroots, J. J. F. (1996). History, concepts, and theory in the psychology of Aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 1-19). San Diego: Academic Press.
- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical assessment. *Lancet, 1*, 307-310.
- Bohnen, N., Twijnstra, A., & Jolles, J. (1992). Performance in the Stroop Color Word Test in relationship to the persistence of symptoms following mild head injury. *Acta Neurologica Scandinavica, 85*, 116-121.
- Bolla, K. I., Gray, S., Resnick, S. M., Galante, R., & Kawas, C. (1998). Category and letter fluency in highly educated older adults. *The Clinical Neuropsychologist, 12*, 330-338.
- Bolla-Wilson, K., & Bleecker, M. L. (1986). Influence of verbal intelligence, sex, age, and education on the Rey auditory verbal learning test. *Developmental Neuropsychology, 2*, 203-211.
- Bosma, H., Van Boxtel, M. P. J., Ponds, R. W. H. M., Houx, P. J., & Jolles, J. (2002). Mental work demands protect against cognitive impairment: MAAS prospective cohort study. *Experimental Aging Research, 29*, 33-45.
- Bosma, H., van Boxtel, M. P. J., Ponds, R. W. H. M., Jelicic, M., Houx, P. J., Metsemakers, J. & Jolles, J. (2000). Engaged lifestyle and cognitive function on middle and old-aged, non-demented persons: a reciprocal association? *Zeitschrift für Gerontologie und Geriatrie, 35*, 575-581.
- Botwinick, J. (1977). Intellectual abilities. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 580-605). New York: Van Nostrand Reinhold.
- Brand, N., & Jolles, J. (1985). Learning and retrieval rate of words presented auditory and visually. *The Journal of General Psychology, 112*, 201-210.
- Brand, N., & Jolles, J. (1987). Information processing in depression and anxiety. *Psychological Medicine, 17*, 145-153.
- Burgess, P. W., Alderman, N., Evans, J., Emslie, H., & Wilson, B. A. (1998). The ecological validity of tests of executive function. *Journal of the International Neuropsychological Society, 4*, 547-558.
- Buschke, H. (1973). Selective reminding analysis of memory and learning. *Journal of Verbal Learning and Verbal Behaviour, 12*, 543-550.
- Cacciatore, F., Abete, P., Ferrare, N., Paolisso, G., Amato, L., Canonico, S., Maggi, S., Varricchio, M., & Rengo, F. (1997). The role of blood pressure in cognitive impairment in an elderly population. *Journal of Hypertension, 15*, 135-142.
- Capitani, E. (1997). Normative data and neuropsychological assessment. Common problems in clinical practice and research. *Neuropsychological Rehabilitation, 7*, 295-309.
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22.
- Chapman, L. J., & Chapman, J. C. (1973). *Disordered thought in schizophrenia*. New York: Appleton-Century-Crofts.
- Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: a review of the literature on everyday cognitive skills. *Neuropsychology Review, 13*, 181-197.
- Chelune, G., Naugle, R.I., Lüders, H., Sedlak, J., & Awad, I.A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology, 7*, 41-52.
- Christensen, H. (2001). What cognitive changes can be expected with normal aging? *Australian and New Zealand Journal of Psychiatry, 35*, 768-775.

References

- Christensen, H., Jorm, A. F., Henderson, A. S., Mackinnon, A. J., Korten, A. E., & Scott, L. R. (1994). The relationship between health and cognitive functioning in a sample of elderly people in the community. *Age and Aging*, 23, 204-212.
- Christensen, H., Korten, A. E., Jorm, A. F., Henderson, A. S., & Scott, L. R. (1996). Activity levels and cognitive functioning in an elderly community sample. *Age and Aging*, 25, 72-80.
- Claus, J., Mohr, E., & Chase, T. (1991). Clinical trials in dementia: Learning effects with repeated testing. *Journal of Psychiatry & Neuroscience*, 16, 1-4.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- Collie, A., Maruff, P., Makdissi, M., McStephen, M., Darby, D. G., & McCrory, P. (2004). Statistical procedures for determining the extent of cognitive change following concussion. *British Journal of Sports Medicine*, 38, 273-278.
- Comalli, P. E., Wapner, S., & Werner, H. (1962). Interference effects of Stroop Color-Word Test in childhood, adulthood, and aging. *Journal of Genetic Psychology*, 100, 47-53.
- Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. London: Chapman and Hall.
- Cowell, P. E., Turetsky, B. I., Gur, R. C., Grossman, R. I., Shtasel, D. L., & Gur, R. E. (1994). Sex differences in aging of the human frontal and temporal lobes. *The Journal of Neuroscience*, 14, 4748-4756.
- Craik, F. I. M., & Salthouse, T. A. (2000). *The handbook of aging and cognition*. Mahwah, New York: Lawrence Erlbaum Associates.
- Crawford, J. R. & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist*, 12, 482-486.
- Crawford, J. R., Obonsawin, M. C., & Bremner, M. (1993). Frontal lobe impairment in schizophrenia: Relationship to intellectual functioning. *Psychological Medicine*, 23, 787-790.
- Crossley, M., D'Arcy, C., & Rawson, N. S. B. (1997). Letter and category fluency in community-dwelling Canadian seniors: A comparison of normal participants to those with dementia of the Alzheimer or vascular type. *Journal of Experimental and Clinical Neuropsychology*, 19, 52-62.
- Crowe, S. F. (1998). Decrease in performance on the Verbal Fluency Test as a function of time: Evaluation in a young healthy sample. *Journal of Clinical and Experimental Neuropsychology*, 20, 391-401.
- Cummings, J. L., Vinters, H. V., Cole, G. M., & Khachaturian, Z. S. (1998). Alzheimer's disease: etiologies, pathophysiology, cognitive reserve, and treatment opportunities. *Neurology*, 51, S2-S17.
- Daigneault, S., Braun, C. M., & Whitaker, H. A. (1992). Early effects of normal aging on perseverative and non-perseverative prefrontal measures. *Developmental Neuropsychology*, 8, 99-114.
- Davidson, D. J., Zacks, R. T., & Williams, C. C. (2003). Stroop Interference, practice and aging. *Aging Neuropsychology and Cognition*, 10, 85-98.
- De Bie, S. E. (1987). *Standaardvragen 1987: Voorstellen voor uniformering van vraagstellingen naar achtergrondkenmerken en interviews* [Standard questions 1987: Proposal for uniformization of questions regarding background variables and interviews]. Leiden, The Netherlands: Leiden University Press.
- De Groot, R. H. M., Hornstra, G., Rozendaal, N., & Jolles, J. (2003). Memory performance, but not information processing speed, may be reduced during early pregnancy. *Journal of Clinical and Experimental Neuropsychology*, 25, 483-489.
- De Ronchi, D., Fratiglioni, L., Rucci, P., Paternico, A., Graziani, S., & Dalmonde, E. (1998). The effect of education on dementia occurrence in an Italian population with middle to high socioeconomic status. *Neurology*, 50, 1231-8.
- Deeg, D. J. H., Knipscheer, C. P. M., & van Tilburg, W. (1993). *Autonomy and well-being in the Aging population: concepts and design of the Longitudinal Aging Study Amsterdam*. Bunnik, The Netherlands: LSOB/NIG.
- Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California Verbal Learning Test Manual*. San Antonio, TX: The Psychological Corporation.
- Dijkstra, J. B., Strik, J. J. M. H., Lousberg, R., Prickaerts, J., Riedel, W. J., Jolles, J., Van Praag, H. M., & Honig, A. (2002). Atypical cognitive profile in patients with depression after myocardial infarction. *Journal of Affective Disorders*, 70, 181-190.
- Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test-retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society*, 5, 346-356.
- Dodrill, C. B. (1997). Myths of neuropsychology. *The Clinical Neuropsychologist*, 11, 1-17.
- Dufouil, C., Alperovitch, A., & Tzourio, C. (2003). Influence of education on the relationship between white matter lesions and cognition. *Neurology*, 60, 831-836.
- Dulaney, C. L., & Rogers, W. A. (1994). Mechanisms underlying reduction in Stroop Interference with practice for young and old adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 470-484.
- Farmer, J. E., & Eakman, A. M. (1995). The relationship between neuropsychological functioning and instrumental activities of daily living following acquired brain injury. *Applied Neuropsychology*, 2, 107-115.
- Fastenau, P. S. (1998). Validity of regression-based norms: an empirical test of the comprehensive norms with older adults. *Journal of Clinical and Experimental Neuropsychology*, 6, 906-916.
- Fastenau, P. S., & Adams, K. M. (1996). Heaton, Grant, and Matthews' Comprehensive Norms: An overzealous attempt. *Journal of Clinical and Experimental Neuropsychology*, 18, 444-448.
- Feinstein, A., Brown, R., & Ron, M. (1994). Effects of practice of serial tests of attention in healthy subjects. *Journal of Clinical and Experimental Neuropsychology*, 16, 436-447.
- Ferraro, F. T., & Balota, D. A. (1999). Memory scanning performance in healthy young adults, healthy older adults, and individuals with dementia of the Alzheimer type. *Aging, Neuropsychology, and Cognition*, 6, 260-272.
- Flowers, K. A., Robertson, C., & Sheridan, M. R. (1995). Some characteristics of word fluency in Parkinson's disease. *Journal of Neurolinguistics*, 9, 33-46.
- Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, 101, 171-191.
- Flynn, J. R. (1999). Searching for justice: the discovery of IQ gains over time. *American Psychologist*, 54, 5-20.

- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-Mental State: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12, 189-198.
- Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.
- Franzen, M. D., & Wilhelm, K. L. (1996). Conceptual foundations of ecological validity in neuropsychology. In R. J. Shorndone & C. J. Long (Eds.), *Ecological Validity of Neuropsychological Testing* (pp. 91-112). Delray Beach, Florida: GR Press/St. Lucie Press.
- Franzen, M. D., Robbins, D. E., & Sawicki, R. F. (1989). *Reliability and validity in neuropsychological assessment*. New York: Plenum press.
- Gaudino, E. A., Geisler, M. W., & Squires, N. K. (1995). Construct validity in the Trail Making Test: what makes part B harder? *Journal of Clinical and Experimental Neuropsychology*, 17(4), 529-535.
- Geffen, G., Moar, K. J., O'Hanlon, A. P., Clark, C. R., & Geffen, L. N. (1990). Performance measures of 16- to 86-year-old males and females on the Auditory Verbal Learning Test. *The Clinical Neuropsychologist*, 4, 45-63.
- Gibbons Natrella, M. (1963). *Experimental Statistics*. New York: Wiley.
- Gleissner, U. & Elger C. (2001). The hippocampal contribution to verbal fluency in patients with temporal lobe epilepsy. *Cortex*, 37, 55-63.
- Golden, C. (1978). *Stroop Color and Word Test: Manual for Clinical and Experimental Uses*. Chicago, IL: Stoelting.
- Graf, P., Uttl, B., & Tuokko, H. (1995). Color- and Picture-word Stroop Tests: performance changes in old age. *Journal of Clinical and Experimental Neuropsychology*, 17, 390-415.
- Greene, R. G. (1992). *Human memory*. Hove, UK: Lawrence Erlbaum Associates Ltd.
- Gulliksen, H. (1973). *Theory of mental tests*. New York: Wiley.
- Haan, M. N., Shemanski, L., Jagust, W. J., Manlino, T. A., & Kuller, L. (1999). The role of APOE epsilon 4 in modulating effects of other risk factors for cognitive decline in elderly persons. *The Journal of the American Medical Association*, 282, 40-46.
- Hameleers, P. A. H. M., Van Boxtel, M. P. J., Hogervorst, E., Riedel, W. J., Houx, P. J., Buntinx, F., & Jolles, J. (2000). Habitual caffeine consumption and its relation to memory, planning capacity and psychomotor performance across multiple age groups. *Human Psychopharmacology: Clinical and Experimental*, 15, 573-581.
- Hammes, J. (1973). *De Stroop Kleur-woord Test: Handleiding [The Stroop Color-Word Test: Manual]*. Amsterdam: Swets & Zeitlinger.
- Harris, M. E., Ivnik, R. J., & Smith, G. E. (2002). Mayo's Older Americans normative studies: expanded AVLT recognition trial norms for ages 57 to 98. *Journal of Clinical and Experimental Neuropsychology*, 24, 214-220.
- Harrison, J. E., Buxton, P., Husain, M., & Wise, R. (2000). Short test of semantic and phonological fluency: Normal performance, validity and test-retest reliability. *British Journal of Clinical Psychology*, 39, 181-191.
- Hays, W. L., & Winkler, R. L. (1971). *Statistics: Probability, Inference, and Decision*. New York: Holt, Rinehart & Winston.
- Hayslip, B., & Panek, P. E. (1989). *Adult development and aging*. New York: Harper & Row.
- Heaton, R. K., Avitable, N., Grant, I., and Matthews, C. G. (1999). Further crossvalidation of regression-based neuropsychological norms with an update for the Boston Naming Test. *Journal of Clinical and Experimental Neuropsychology*, 21, 572-582.
- Heaton, R. K., Matthews, C. G., Grant, I., & Avitable, N. (1996). Demographic corrections with comprehensive norms: An overzealous attempt, or a good start? *Journal of Clinical and Experimental Neuropsychology*, 18, 449-458.
- Heaton, R. K., Temkin, N., Dikmen, S., Avitable, N., Taylor, M. J., Marcotte, T. D., et al. (2001). Detecting change: a comparison of three neuropsychological methods, using normal and clinical samples. *Archives of Clinical Neuropsychology*, 16, 75-91.
- Henry, J. D., & Crawford, J. R. (2004). A Meta-analytic review of verbal fluency performance following focal cortical lesions. *Neuropsychology*, 18, 284-295.
- Herlitz, A., & Yonker, J. E. (2002). Sex differences in episodic memory: The influence of intelligence. *Journal of Clinical and Experimental Neuropsychology*, 24, 107-114.
- Hertzog, C. (1996). Research design in studies of aging and cognition. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 1-19). San Diego: Academic Press.
- Heubrock, D. (1994). Auditiv-verbales Lernen unter standardisierten Bedingungen: Erste deutsche Normen für 16- bis 26-jährige Männer und Frauen zum Auditiv-Verbalen Lerntest (AVLT) [Standardized auditory-verbal learning: Preliminary German Auditory-Verbal Learning Test (AVLT) norms for a male and female population aged 18 to 26]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 15, 65-76.
- Hijman, R., Jolles, J., Verhoeven, W.M.A., Van Ree, J.M., Elderson, A., & De Wied, D. (1992). Desglycinamide-(Arg8)-vasopressin in five trials with memory-disturbed patients. *Human Psychopharmacology: Clinical and Experimental*, 7, 7-23.
- Horn, J. L. (1982). The aging of human abilities. In B. B. Wolman (Ed.), *Handbook of developmental psychology* (pp. 847-870). New York: Wiley.
- Houx, P. J. (1991). *Cognitive aging and health-related factors*. Maastricht, the Netherlands: Neuropsych publishers.
- Houx, P. J., Jolles, J., & Vreeling, F. W. (1993). Stroop Interference: Aging effects associated with the Stroop Color-Word Test. *Experimental Aging Research*, 19, 209-224.
- Houx, P. J., Shepherd, J., Blauw, G. J., Murphy, M. B., Ford, I., Bollen, E. L. et al. (2002). Testing cognitive function in elderly populations: the PROSPER Study. *Journal of Neurology Neurosurgery and Psychiatry*, 73, 385-389.
- Hultsch, D. F., Hertzog, C., Dixon, R. A., & Small, B. J. (1998). *Memory change in the aged*. New York: Cambridge University Press.

- Hurks, P. P. M., Vles, J. S. H., Hendriksen, J. G. M., Kalff, A. C., Feron, F. J. M., Kroes, M., Van Zeven, T. M. C., Steyaert, J., & Jolles, J. (2004). Semantic category fluency versus initial letter fluency over time as a measure of automatic and controlled processing in healthy school-aged children. *Brain and Cognition*, 55, 535-544.
- Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., & Petersen, R. C. (1996). Neuropsychological tests' norms above age 55: COWAT, BNT, MAE Token, WRAT-R Reading, AMNART, STROOP, TMT, and JLO. *The Clinical Neuropsychologist*, 10, 262-278.
- Ivnik, R. J., Smith, G. E., Lucas, J. A., Petersen, R. C., Boeve, B. F., Kokmen, E., (1999). Testing normal older people three or four times at 1- to 2-year intervals: Defining normal variance. *Neuropsychology*, 13, 121-127.
- Jacobson, N. S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12-19.
- Jarvis, P. E., & Barth, J. T. (1987). *Halstead-Reitan test battery: An interpretive guide*. Odessa, FL: Psychological Assessment Resources.
- Jelicic, M., Van Boxtel, M. P. J., Houx, P. J., & Jolles, J. (2000). Does migraine headache affect cognitive function in the elderly? Report from the Maastricht Aging Study (MAAS). *Headache*, 40, 715-719.
- Jolles, J., Houx, P. J., Van Boxtel, M. P. J., & Ponds, R. W. H. M. (1995). *Maastricht Aging Study: Determinants of Cognitive Aging*. Maastricht, The Netherlands: Neuropsych Publishers.
- Jorm, A. F., & Jolley, D. (1998). The incidence of dementia: A meta-analysis. *Neurology*, 51, 728-733.
- Kempler, D., Teng, E. L., Dick, M., Taussig, M., Davis, D. S. (1998). The effects of age, education, and ethnicity on verbal fluency. *Journal of the International Neuropsychological Society*, 4, 531-538.
- Kibby, M. Y., Schmitter-Edgecombe, M., & Long, C. J. (1998). Ecological validity of neuropsychological tests: focus on the California Verbal Learning Test and the Wisconsin Card Sorting Test. *Archives of Clinical Neuropsychology*, 13, 523-534.
- Klein, M., Ponds, R. W., Houx, P. J., & Jolles, J. (1997). Effect of test duration on age-related differences in Stroop interference. *Journal of Clinical and Experimental Neuropsychology*, 19, 77-82.
- Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariate methods* (third edition). New York: Duxbury Press.
- Kosmidis, M. H., Vlahou, C. H., Panagiotaki, P., & Kiosseoglou, G. (2004). The verbal fluency task in the Greek population: Normative data, and clustering and switching strategies. *Journal of the International Neuropsychological Society*, 10, 164-172.
- Krabbedam, L., Van Harten, P. N., Picus, I., & Jolles, J. (2000). Tardive dyskinesia is associated with impaired retrieval from long-term memory: the Curacao extrapyramidal syndromes study: IV. *Schizophrenia Research*, 42, 41-46.
- La Rue, A. (1992). *Aging and Neuropsychological Assessment*. New York: Plenum Press.
- Lamberts, H., & Wood, M. (1987). *ICPC. International Classification of Primary Care*. Oxford: Oxford University Press.
- Lannoo, E., & Vingerhoets, G. (1997). Flemish normative data on common neuropsychological tests: Influence of age, education, and sex. *Psychologica Belgica*, 37, 141-155.
- LaRue, A. (1992). *Aging and neuropsychological assessment*. New York: Plenum Press.
- Le Carret, N., Lafont, S., Letenneur, L., Dartigues, J. F., Mayo, W., & Fabrigoule, C. (2003). The effect of education on cognitive performances and its implication for the constitution of the cognitive reserve. *Developmental Neuropsychology*, 23, 317-337.
- LeBlanc, J. M., Hayden, M. E., & Paulman, R. G. (2000). A comparison of neuropsychological assessment for predicting employability after closed head injury. *Journal of Head Trauma Rehabilitation*, 15, 1022-1040.
- Lee, D. Y., Lee, K. U., Lee, J. H., Kim, K. W., Jhoo, J. H., Kim, S. Y., Yoon, J. C., Woo, S. I., Ha, J., & Woo, J. I. (2004). A normative study of the CERAD neuropsychological assessment battery in Korean elderly. *Journal of the International Neuropsychological Society*, 10, 72-81.
- Lezak, M. D. (1995). *Neuropsychological Assessment* (3rd ed.). New York: Oxford University Press.
- Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological Assessment* (fourth edition). New York: Oxford University Press.
- Libon, D. J., Glosser, G., Malamut, B. L., Kaplan, E., Goldberg, E., Swenson, R., & Sands, L. P. (1994). Age, executive functions, and visuospatial functioning in healthy older adults. *Neuropsychology*, 8, 38-43.
- Linschoten, J. (1963). *De la Court frequentie tellingen van Nederlandse woorden* (Report No. 6301). Rijks Universiteit Utrecht, Psychologisch Laboratorium.
- Loewenstein, D. A., & Rubert, M. P. (1992). The NINCDS-ADRDA neuropsychological criteria for the assessment of dementia: Limitations of current diagnostic guidelines. *Behavior, Health and Aging*, 2, 113-121.
- Loewenstein, D. A., Arguelles, T., Arguelles, S., & Linn-Fuentes, P. (1994). Potential cultural bias in the neuropsychological assessment of the older adult. *Journal of Clinical and Experimental Neuropsychology*, 16, 623-629.
- Logie, H., Hogan, R., & Peut, A. (2004). *Longitudinal studies of ageing: Implications for future studies*. Canberra: AIHW.
- Lopez, S. R., & Taussig, F. M. (1991). Cognitive-intellectual functioning of Spanish-speaking impaired and nonimpaired elderly: Implications for culturally sensitive assessment. *Psychological Assessment*, 3, 448-454.
- Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 304-305.
- Lord, F. M. (1969). Statistical adjustments when comparing pre-existing groups. *Psychological Bulletin*, 72, 336-337.
- Lord, F., & Novick, M. (1968). *Statistical theories of mental tests*. New York: Addison-Wesley.
- Luteijn, F., & Van der Ploeg, F. A. E. (1983). *Handleiding Groninger Intelligentietest (GIT)* [Manual Groningen Intelligence Test]. Lisse, The Netherlands: Swets and Zeitlinger.
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19-40.

- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109, 163-203.
- Marquardt, D. W. (1980). You should standardize the predictor variables in your regression models. *Journal of the American Statistical Association*, 75, 87-91.
- Martin, N. J., & Franzen, M. D. (1989). The effect of anxiety on neuropsychological function. *International Journal of Clinical Neuropsychology*, 11, 1-8.
- Mathuranath, P. S., George, A., Cherian, P. J., Alexander, A., Sarma, S. G., & Sarma, P. S. (2003). Effects of age, education and gender of verbal fluency. *Journal of Clinical and Experimental Neuropsychology*, 25, 1057-1064.
- McCaffrey, R., Ortega, A., & Haase, R. (1993). Effects of repeated neuropsychological assessments. *Archives of Clinical Neuropsychology*, 8, 519-524.
- Meijer, W. A., Van Boxtel, M. P. J., Van Gerven, P., Van Hooren, S., De Groot, R. H. M., & Jolles, J. (2006). *Interaction effects of education and health status on cognitive change: 6-year follow-up of the Maastricht Aging Study*. Manuscript submitted for publication.
- Meinz, E. J., & Salthouse, T. A. (1998). Is age kinder to females than to males? *Psychonomic Bulletin & Review*, 5, 56-70.
- Metsemakers, J. F. M., Höppener, P., Knottnerus, J. A., Kocken, R. J. J., & Limonard, C. B. G. (1992). Computerized health information in the Netherlands: A registration network of family practices. *British Journal of General Practice*, 42, 102-106.
- Miner, T., & Ferraro, F. R. (1998). The role of speed of processing, inhibitory mechanisms, and presentation order in Trail-Making Test performance. *Brain and Cognition*, 38, 246-253.
- Mitrushina, M. N., Boon, K. B., & D'Elia, L. F. (1999). *Handbook of Normative Data for Neuropsychological Assessment*. New York: Oxford University Press.
- Moering, R. G., Schinka, J. A., Mortimer, J. A., & Graves, A. B. (2003). Normative data for elderly African Americans for the Stroop Color and Word Test. *Archives of Clinical Neuropsychology*, 607, 1-11.
- Mol, M. E. M., van Boxtel, M. P. J., Willems, D., & Jolles, J. (2006). Do subjective memory complaints predict cognitive dysfunction over time? A six-year follow-up of the Maastricht Aging Study. *International Journal of Geriatric Psychiatry*, 21, 432-441.
- Møller, J. T., Cluitmans, P., Rasmussen, L. S., Houx, P. J., Rasmussen, H., Canet, J., Rabbitt, P., Jolles, J., Larsen, K., Han-ning, C. D., Langeron, O., Johnson, T., Lauven, P. M., Kristensen, P. A., Biedler, A., van Beem, H., Fraidakis, O., Silverstein, J. H., Beneken, J. E. W., & Gravenstein, J. S. (1998). Long-term postoperative cognitive dysfunction in the elderly: ISPOCD1 study. *The Lancet*, 351, 857-861.
- Mungas, D., & Reed, B. R. (2000). Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Statistics in Medicine*, 19, 1631-1644.
- Mungas, D., Reed, B. R., & Kramer, J. H. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology*, 17, 380-392.
- Murphy, D. G. M., DeCarli, C., McIntosh, A. R., Daly, E., Mentis, M. J., Pietrini, P., (1996). Sex differences in human brain morphometry and metabolism: and in vivo quantitative magnetic resonance imaging and positron emission tomography study on the effects of aging. *Archives of General Psychiatry*, 53, 585-594.
- Naus, M. J., Cermak, L. S., & DeLuca, D. (1977). Retrieval processes in alcoholic Korsakoff patients. *Neuropsychologia*, 5, 737-742.
- Neter, J., Wasserman, W., & Kutner, M. H. (1989). *Applied linear regression models* (2nd ed.). Homewood, IL: Irwin.
- Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. *Neuropsychologia*, 9, 97-113.
- Pachana, N. A., Boone, K. B., Miller, B. L., Cummings, J. L., & Berman, N. (1996). Comparison of neuropsychological functioning in Alzheimer's disease and frontotemporal dementia. *Journal of the International Neuropsychological Society*, 2, 505-510.
- Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory and Cognition*, 17, 398-422.
- 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.
- Poewe, W., Berger, W., Benke, T., & Schelosky, L. (1991). High-speed memory scanning in Parkinson's disease: Adverse effects of Levodopa. *Annals of Neurology*, 29, 670-673.
- Ponds, R. W. H. M., Van Boxtel, M. P. J., & Jolles, J. (2000). Age-related changes in subjective cognitive functioning. *Educational Gerontology*, 26, 67-81.
- Pontius, A. A., & Yudowitz, B. S. (1980). Frontal lobe system dysfunction in some criminal actions as shown in the narratives test. *Journal of Nervous and Mental Disease*, 168, 111-117.
- Raskin, S. A., & Rearick, E. (1996). Verbal fluency in individuals with mild traumatic brain injury. *Neuropsychology*, 10, 416-422.
- Rasmussen, L. S., Siersma V. D., & the ISPOCD group (2004). Postoperative cognitive dysfunction: true deterioration versus random variation. *Acta Anaesthesiologica Scandinavica*, 48, 1137-1143.
- Rasquin, S. M. C., Lodder, J., Ponds, R. W. H. M., Winkens, I., Jolles, J., & Verhey, F. R. J. (2004). Cognitive functioning after stroke: a one-year follow-up study. *Dementia and Geriatric Cognitive Disorders*, 18, 138-144.
- Ratcliff, G., Ganguli, M., Chandra, V., Sharma, S., Belle, S., Seaberg, E., & Pandav, R. (1998). Effects of literacy on measures of word fluency. *Brain and Language*, 61, 115-122.
- Rey, A. (1958). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.
- Rideout, B. E., & Winchester, L. (1990). Effects on verbal fluency associated with sex, practice, and stimulus letter. *Perceptual and Motor Skills*, 71, 656-658.

- Ruff, R. M., Light, R. H., Parker, S. B., & Levin, H. S. (1996). Benton Controlled Oral Word Association Test: Reliability and updated norms. *Archives of Clinical Neuropsychology*, 11, 329-338.
- Ruff, R. M., Light, R. H., Parker, S. B., & Levin, H. S. (1997). The psychological construct of word fluency. *Brain and Language*, 57, 394-405.
- Salthouse, T. A. (1992). Why do adult age differences increase with task complexity? *Developmental Psychology*, 28, 905-918.
- Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, 103, 403-428.
- Salthouse, T. A., Toth, J., Daniels, K., Parks, C., Pak, R., Wolbrette, M., et al. (2000). Effects of aging on efficiency of task switching in a variant of the Trail Making Test. *Neuropsychology*, 14, 102-111.
- Satz, P. (1993). Brain reserve capacity on symptom onset after brain injury: A formulation and review of evidence for threshold theory. *Neuropsychology*, 7, 273-295.
- Savage, R. M., & Gouvier, W. D. (1992). Rey auditory-verbal learning test: the effects of age and gender, and norms for delayed recall and story recognition trials. *Archives of Clinical Neuropsychology*, 7, 407-414.
- Sbordone, R. J. (1996). Ecological validity: some critical issues for the neuropsychologist. In: R. J. Sbordone & C. J. Long (Eds.), *Ecological Validity of Neuropsychological Testing* (pp. 15-42). Delray Beach, Florida: GR Press/St. Lucie Press.
- Scarmeas, N., & Stern, N. (2004). Cognitive reserve: Implications for diagnosis and prevention of Alzheimer's disease. *Current Neurology and Neuroscience Reports*, 4, 374-380.
- Scarmeas, N., & Stern, Y. (2003). Cognitive reserve and lifestyle. *Journal of Clinical and Experimental Neuropsychology*, 25, 625-633.
- Scarmeas, N., Levy, G., Tang, M. X., Manly, J., & Stern, Y. (2001). Influence of leisure activity on the incidence of Alzheimer's Disease. *Neurology*, 57, 2236-2242.
- Schaie, K. W. (1988). Internal validity threads in studies of adult cognitive development. In M. L. Lowe & C. J. Brainerd (Eds.), *Cognitive development in adulthood*. New York: Springer Verlag.
- Schaie, K. W. (1996). Intellectual development in adulthood. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 266-286). San Diego: Academic Press.
- Schmand, B., Smit, J. H., Geerlings, M. I., & Lindeboom, J. (1997b). The effects of intelligence and education on the development of dementia. A test of the brain reserve hypothesis. *Psychological Medicine*, 27, 1337-1344.
- Schmand, B., Smit, J., Lindeboom, J., Smits, C., Hooijer, C., Jonker, C., & Deelman, B. (1997a). Low education is a genuine risk factor for accelerated memory decline and dementia. *Journal of Clinical Epidemiology*, 50, 1025-1033.
- Schmidt, M. (1996). *Rey Auditory Verbal Learning Test: A Handbook*. Los Angeles: Western Psychological Services.
- Schmitt, J. A. J., Jorissen, B. L., Sobczak, S., Van Boxtel, M. P. J., Hogervorst, E., Deutz, N. E. P., & Riedel, W. J. (2000). Tryptophan depletion impairs memory consolidation but improves focused attention in healthy young volunteers. *Journal of Psychopharmacology*, 14, 21-29.
- Sergeant, J. A., Geurts, H., & Oosterlaan, J. (2002). How specific is a deficit of executive functioning for Attention-Deficit/Hyperactivity Disorder? *Behavioural Brain Research*, 130, 3-28.
- Smith, A. (1982). *Symbol Digit Modalities Test, manual*. Los Angeles, CA: Western Psychological Services.
- Spreen, O., & Strauss, E. (1998). *A compendium of Neuropsychological Tests*. New York: Oxford University Press.
- Squire, L. R. (1987). *Memory and brain*. London: Oxford University Press.
- Stern, R. A., Silva, S. G., Chaisson, N., & Evans, D. L. (1996). Influence of cognitive reserve on neuropsychological functioning in asymptomatic human immunodeficiency virus-1 infection. *Archives of Neurology*, 53, 148-153.
- Stern, Y. (2003). The concept of cognitive reserve: A catalyst for research. *Journal of Clinical and Experimental Neuropsychology*, 25, 589-593.
- Stern, Y., Zarahn, E., Hilton, H. J., Flynn, J., DeLaPaz, R., & Rakitin, B. (2003). Exploring the neural basis of cognitive reserve. *Journal of Clinical and Experimental Neuropsychology*, 25, 691-701.
- Sternberg, S. (1966). High-speed memory scanning in human memory. *Science*, 153, 652-654.
- Sternberg, S. (1975). Memory scanning: New findings and current controversies. *Quarterly Journal of Experimental Psychology*, 27, 1-32.
- Storandt, M., Botwinick, J., Danziger, W. L., Berg, L., & Hughes, C. P. (1984). Psychometric differentiation of mild senile dementia of the Alzheimer type. *Archives of Neurology*, 41, 497-499.
- Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643-662.
- Swerdlow, N. R., Filion, D., Geyer, M. A., & Braff, D. L. (1995). "Normal" personality correlates of sensorimotor, cognitive, and visuospatial gating. *Biological Psychiatry*, 37, 286-299.
- Taub, H. A. (1972). A comparison of young adult and old groups on various digit span tasks. *Developmental Neuropsychology*, 6, 60-65.
- Taub, H. A. (1973). Memory span, practice, and aging. *Journal of Gerontology*, 28, 335-338.
- Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society*, 5, 357-369.
- Thompson, J. C., Stopford, C. L., Snowden, J. S., & Neary, D. (2005). Qualitative neuropsychological performance characteristics in frontotemporal dementia and Alzheimer's disease. *Journal of Neurology, Neurosurgery, & Psychiatry*, 76, 920-927.
- Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, University of Chicago Press.
- Tierney, M. C., Blach, S. E., Szalai, J. P., Snow, G., Fisher, R. H., Nadon, G., & Chui, H. C. (2001). Recognition memory and verbal fluency differentiate probable Alzheimer disease from subcortical ischemic vascular dementia. *Archives of Neurology*, 58, 1654-1659.

- Tombaugh, T. N., Kozak, J., Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Clinical Neuropsychology*, 14, 167-177.
- Tomer, R., & Levin, B. E. (1993). Differential effects of aging on two verbal fluency tasks. *Perceptual and Motor Skills*, 76, 465-466.
- Trenerry, M., Crosson, B., DeBoe, J., & Leber, W. (1989). *Stroop Neuropsychological Screening Test, Manual*. Adessa, FL: Psychological Assessment Resources (PAR).
- Uchiyama, C. L., D'Elia, L. F., Dellinger, A. M., Becker, J. T., Selnes, O. A., Welsch, J. E., Chen, B. B., Satz, P., Van Gorp, W., & Miller, E. N. (1995). Alternate forms of the auditory-verbal learning test: Issues of test comparability, longitudinal reliability, and moderating demographic variables. *Archives of Clinical Neuropsychology*, 10, 133-145.
- United Nations Educational, Scientific and Cultural Organisation (UNESCO) (1976). *International Standard Classification of Education (ISCED)*, UNESCO, Paris.
- Uttl, B., & Graf, P. (1997). Color-Word Stroop test performance across the adult life span. *Journal of Clinical and Experimental Neuropsychology*, 19, 405-420.
- Valentijn, S. A. M., van Boxtel, M. P. J., van Hooren, S. A. H., Bosma, H., Beckers, H. J. M., Ponds, R. W. H. M., et al. (2005). Change in Sensory Functioning Predicts Change in Cognitive Functioning: Results from a 6-Year Follow-Up in the Maastricht Aging Study. *Journal of the American Geriatrics Society*, 53(3), 374-380.
- Van Beijsterveldt, C. E. M., van Boxtel, M. P. J., Bosma, H., Houx, P. J., Buntinx, F., & Jolles, J. (2002). Predictors of attrition in a longitudinal cognitive aging study: The Maastricht Aging Study (MAAS). *Journal of Clinical Epidemiology*, 55, 216-223.
- Van Boxtel, M. P. J., Buntinx, F., Houx, P. J., Metsemakers, J. F. M., Knottnerus, A., & Jolles, J. (1998). The relation between morbidity and cognitive performance in a normal aging population. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 53A, M147-M154.
- Van Boxtel, M. P. J., Paas, F. G. W. C., Houx, P. J., Adam, J. J., Teeken, J. C., & Jolles, J. (1997). Aerobic capacity and cognitive performance in a cross-sectional aging study. *Epidemiology*, 29, 1357-1365.
- Van Boxtel, M. P. J., ten Tusscher, M. P. M., Metsemakers, J. F. M., Willems, B., & Jolles, J. (2001). Visual determinants of reduced performance on the Stroop Color-Word Test in normal aging individuals. *Journal of Clinical and Experimental Neuropsychology*, 23, 620-627.
- Van Boxtel, M. P. J., van Beijsterveldt, C. E. M., Houx, P. J., Anteunis, L. J. C., Metsemakers, J. F. M., & Jolles, J. (2000). Mild hearing impairment can reduce verbal memory performance in a healthy adult population. *Journal of Clinical and Experimental Neuropsychology*, 22, 147-154.
- Van Boxtel, P. J., Gaillard, C., Houx, P. J., Buntinx, F., de Leeuw, P. W., & Jolles, J. (1997). Can blood pressure predict cognitive task performance in a healthy population sample? *Journal of Hypertension*, 15, 1069-1076.
- Van Breukelen, G. J. P. (2006). ANCOVA vs. change from baseline: More power in randomized studies, more bias in non-randomized studies. *Journal of Clinical Epidemiology*, 59, in press.
- Van Breukelen, G. J. P., & Vlaeyen, J. W. S. (2005). Norming clinical questionnaires with multiple regression: The Pain Cognition List. *Psychological Assessment*, 17, 336-344.
- Van den Burg, W., & Kingma, A. (1999). Performance of 225 Dutch school children on Rey's Auditory Verbal Learning Test (AVLT): Parallel test-retest reliabilities with an interval of 3 months and normative data. *Archives of Clinical Neuropsychology*, 14, 545-559.
- Van der Elst, W., van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2005). Rey's Verbal Learning Test: Normative data for 1855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society*, 11, 290-302.
- Van der Elst, W., van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006a). The Stroop Color-Word Test: influence of age, sex, and education; and normative data for a large sample across the adult age range. *Assessment*, 13, 62-79.
- Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006b). The Letter Digit Substitution Test: Normative data for 1,858 healthy participants aged 24-81 from the Maastricht Aging Study (MAAS): Influence of age, education, and sex. *Journal of Clinical and Experimental Neuropsychology*, 28, 998-1009.
- Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006c). Normative data for the animal, profession and letter M naming verbal fluency tests for Dutch speaking participants and the effects of age, education, and sex. *Journal of the International Neuropsychological Society*, 12, 80-89.
- Van Hoof, J. J. M., Jogems-Kosterman, B. J. M., Sabbe, B. G. C., Zitman, F. G., & Hulstijn, W. (1998). Differentiation of cognitive and motor slowing in the Digit Symbol Test (DST): differences between depression and schizophrenia. *Journal of Psychiatric Research*, 32, 99-103.
- Van Hooren, S. A., Anteunis, L. J., Valentijn, S. A., Bosma, H., Ponds, R. W., Jolles, J., & van Boxtel, M. P. (2005). Does cognitive function in older adults with hearing impairment improve by hearing aid use? *International Journal of Audiology*, 44, 265-271.
- Videbech, P., Ravnkilde, B., Kristensen, S., Egander, A., Clemmensen, K., Rasmussen, N. A., Gjedde, A., & Rosenberg, R. (2003). The Danish PET/depression project: poor verbal fluency performance despite normal prefrontal activation in patients with major depression. *Psychiatry Research: Neuroimaging*, 123, 49-63.
- Vink, M., & Jolles, J. (1985). A new version of the Trail Making Test as an information processing task. *Journal of Clinical Neuropsychology*, 7, 162.
- Ware, J. J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form healthy survey (SF-36): I. conceptual framework and item selection. *Medical Care*, 30, 473-483.
- Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corporation.
- Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-Revised*. New York: Psychological Corporation.
- West, R. L. (1996). An application of prefrontal cortex function theory to cognitive aging. *Psychological Bulletin*, 120, 272-292.

References

- Whitfield, K. E., Seeman, T. E., Miles, T. P., Albert, M. S., Berkman, L. F., & Blazer, D. G. Health indices as predictors of cognition among African Americans: MacArthur studies of successful aging. *Ethnicity & Disease*, 7, 127-136.
- Wickens, C. D., Braune, R., & Stokes, A. (1987). Age differences in the speed and capacity of information processing: 1. A dual-task approach. *Psychology & Aging*, 2, 70-80.
- Williams, J. D., & Klug, M. G. (1996). Aging and cognition: Methodological differences in outcome. *Experimental Aging Research*, 22, 219-244.
- Wilson, B. A., Alderman, N., Burgess, P., Emslie, H., & Evans, J. J. (1996). *Behavioural Assessment of the Dysexecutive Syndrome (BADS)*. Thames Valley Test Company, Bury St. Edmunds, England.
- Woodruff-Pak, D. S. (1997). *The Neuropsychology of Aging*. Malden: Blackwell Publishers.
- Woods, S. P., Conover, E., Rippeth, J. D., Carey, C. L., Gonzalez, R., Marcotte, T. D., Heaton, R. K., & Grant, I. (2004). Qualitative aspects of verbal fluency in HIV-associated dementia: A deficit in rule-guided lexical-semantic search processes? *Neuropsychologia*, 42, 801-809.
- Yonker, J. E., Eriksson, E., Nilsson, L., & Herlitz A. (2003). Sex differences in episodic memory: Minimal influence of estradiol. *Brain and Cognition*, 52, 231-238.
- Zachary, R. A., & Gorsuch, R. L. (1985). Continuous norming: implications for the WAIS-R. *Journal of Clinical Psychology*, 41, 86-97.

Summary

The number of older people in Western societies increases rapidly. Consequently, it is important to gain more insight in the process of aging. The present thesis focussed on the cognitive aspects of normal aging, from a theoretical and an applied viewpoint. The relevance and relative contribution of age and age-extrinsic factors on inter-individual differences in performance on cognitive tests was formally modelled, as well as the contribution of these factors on the changes in cognitive performance over time. These formal models increase the insight in the patterns, onset, and magnitude of age-related differences and changes in cognitive abilities. An applied aim of the present thesis was to establish the normal range of performance for six neuropsychological tests that measure cognitive abilities that are typically assessed in most neuropsychological examinations. Such norms are a fundamental tool for clinicians and researchers, because they refer to normative data whenever they evaluate the cognitive abilities of an individual or a patient, for example when evaluating cognitive impairment in people who are suspected of conditions such as dementia.

In Chapter 1, the aims and approach of this thesis are described in more detail. Existing knowledge on cognitive aging is reviewed, and background information regarding neuropsychological assessment, neuro-psychometrics, and the Maastricht Aging Study (MAAS) is provided.

In Chapter 2, the effect of age, gender, and level of education on performance on the Stroop Color Word Test (SCWT) was evaluated, and the normal range of performance was established. The SCWT measures executive functioning and related concepts such as attention. The analyses were based on a MAAS sample of 1,856 healthy participants aged between 24 and 81. Multiple regression analyses showed that the speed-dependent SCWT scores were profoundly affected by the demographic variables, rather than the accuracy measures (i.e. errors). The age-related decline in executive functions was suggested to be more pronounced for lower educated people as compared to their higher educated counterparts, a result that is in line with the cognitive reserve theory. Traditional and regression-based normative data were provided. A comparison between both methods suggested that the regression-based normative method is superior to the traditional method, although the latter approach is the most frequently used normative method until now.

Chapter 3 describes the results of a study that evaluated the effects of age and age-extrinsic variables on the performance on the Verbal Learning Test of Rey (VLT). The VLT is one of the most commonly used neuropsychological tests for declarative memory. The words to be learned were presented auditorily or visually in MAAS. The effect of mode of presentation on test performance was evaluated in addition to the effects of the demographical variables because previous research suggested that auditorily presented material is recalled better

than visual material on short-term memory tasks (the “modality effect”). The VLT scores of a sample of 1,855 healthy participants aged between 24 and 81 from the MAAS study were regressed on mode of presentation, age, gender, and level of education. The results showed that VLT performance decreased from an early age, and that females and higher educated people performed better than males and lower educated people. The effect of presentation mode on VLT performance was mixed. The pattern of results suggested that auditory presentation was associated with better performance on VLT trials that depend on short-term memory processes, whilst visual presentation was associated with a better performance of tasks that depend on long-term memory processes or tasks that depend on a mixture of short-term and long-term memory processes. Regression-based VLT norms were provided, corrected for the relevant independent variables.

The aim of Chapter 4 was to determine whether age, gender, and level of education affected performance on semantic and phonemic Verbal Fluency Tests (VFTs). The Animal, Profession, and Letter *M* naming VFTs were administered to a sample of 1,856 people who were aged between 24 and 81 years. Regression analyses showed that age affected performance on all VFTs profoundly, but that the age-effect was not uniform. For the Profession and the Letter *M* naming VFTs, performance remained quite stable in young adulthood but declined strongly after age 50. In the Animal naming VFT on the other hand, performance declined linearly, starting early in life. Gender only affected the performance on the Profession naming VFT, and Education was found to affect the performance on all VFTs profoundly. Regression-based norms were provided, appropriately corrected for the relevant demographical variables.

In Chapter 5, the Concept Shifting Test (CST) was described. The CST is a modification of the well-known Trail Making Test (TMT). The CST was especially developed to avoid certain problems with the TMT that hamper specificity of the executive function measure. A total sample of 1,856 participants aged between 24 and 81 did the CST in MAAS. Based on this sample, the influence of age, gender, level of education and handedness on CST performance was determined by using multiple linear and (binary) logistic regression analyses. The results showed that CST performance was superior in women, decreased with age, and increased with level of education. Normative data that are corrected for the influences of these demographical variables were established using a regression-based normative approach.

Chapter 6 presented a paper-and-pencil version of Sternberg’s Memory Scanning Test (P&P MST). This new test was especially developed for use in the clinic, because the original Sternberg task is difficult to use in a clinical environment. For example, the administration time of the original MST is long and the test is too difficult for old people or people with cognitive disorders. As is the case for the MST, the P&P MST is used to measure memory scanning and non-memory functions in working memory. A total sample of 1,839 people aged between 24 and 81 were administered the P&P MST. Age and education

affected all components of information processing in working memory. Gender only affected the speed of memory scanning. The results suggested that younger females (aged below 55) scanned working memory faster than males, and vice versa for older people (aged above 55). Normative data were provided for the non-memory processing and the memory scanning P&P MST scores, stratified by the relevant independent variables.

Chapter 7 evaluated the influence of demographical variables on performance on the written and oral versions of the Letter Digit Substitution Test (LDST). This test is an adaptation of the procedurally identical Digit Symbol Substitution Test. The LDST provides a general measure of speed of information processing. The written and oral versions of the LDST were administered to a sample of 1,858 adults aged 24 to 81 years in MAAS. Age, gender, and level of education affected LDST performance profoundly. Regression-based normative data were prepared for both the written and oral versions of the LDST.

In Chapter 8, the influence of demographical variables on the change in performance after three years in the visual version of the VLT, the SCWT and the LDST was evaluated. A large sample of people aged between 49 and 81 at baseline were tested twice with a test-retest interval of 3 years. The change scores were regressed on age, gender, and level of education. Test-retest changes differed as a function of age and/or gender for the SCWT Interference, VLT Total recall 1-3, VLT Delayed recall, and LDST measures, but were identical across subgroups for the VLT Immediate recall measure. The results also suggested that the age-related decline in executive functioning after three years was more pronounced for males than for females. The normal range of change in performance was established, with appropriate corrections for the relevant independent variables.

In Chapter 9, concluding remarks are presented about 1) cognitive reserve and aging, 2) regression-based versus traditional normative procedures, 3) viable independent variables in normative research, and 4) ecological validity and neuropsychological assessment. Finally, some suggestions for improving the methods and approaches of neuropsychological assessment in the future are given. On the basis of the results that were presented in this thesis, it can be concluded that age negatively influences most, if not all, fluid cognitive abilities. For most cognitive measures, linear as well as quadratic age-effects were observed. It was also shown that factors such as gender and education should be taken into account when inter-individual differences in cognitive abilities are evaluated or studied. The influence of the various demographical variables on cognitive functioning was shown to differ across cognitive domains. Formal modeling by using multiple linear regression analysis (or logistic regression analysis) provides an elegant approach to evaluate the relevance and relative contribution of multiple independent variables on inter-individual differences in cognition and on change in cognition over time. When it comes to establishing the normal range of performance on (neuro)psychological tests, the traditional

normative method should be abandoned in favor of the regression-based method. Indeed, in contrast to the traditional normative method, the regression-based approach (i) provides more accurate estimates of the population parameters, (ii) avoids the problem of unbalanced data in normative samples, (iii) can handle continuous, categorical and dummy variables as predictors of test performance, (iv) allows for correcting the norms for a large number of independent variables, and (v) provides a general framework for establishing norms that applies to various situations. The regression-based norms that were established in the present thesis provide an important tool for many neuropsychologists and researchers in the Netherlands and abroad.

Samenvatting

Het aantal ouderen in de Westerse samenleving stijgt snel. Het is daarom van belang om een goed inzicht te verwerven in het verouderingsproces. Dit proefschrift focust vanuit een theoretisch en toegepast perspectief op de cognitieve aspecten van de normale veroudering. De relevantie en relatieve bijdrage van leeftijd en leeftijdsextrinsieke factoren op inter-individuele variatie in de prestatie op cognitieve taken werd formeel statistisch gemodelleerd, alsook de bijdrage van deze variabelen aan de verandering in cognitieve prestaties die optreden als functie van tijd. Zulke formele modellen hebben tot doel het inzicht in de patronen, aanvang en grootte van leeftijdsgerelateerde verschillen en veranderingen in cognitieve functies te vergroten. Een toegepast doel van dit proefschrift was om de normale variabiliteit in prestatie op zes neuropsychologische taken te bepalen. Normen voor zulke taken vormen een fundamenteel werkinstrument voor klinici en onderzoekers die het cognitief functioneren van een persoon evalueren, bijvoorbeeld in het kader van het diagnosticeren van aandoeningen zoals dementie.

In Hoofdstuk 1 werden de doelen en onderzoeksmethoden die in dit proefschrift werden aangewend meer gedetailleerd beschreven. Reeds bestaande kennis over de cognitieve veroudering werd samengevat, en achtergrond informatie over neuropsychologische assessment, neuro-psychometrie, en de Maas-tricht Aging Study (MAAS) werd gegeven.

In Hoofdstuk 2 werd het effect van leeftijd, geslacht, en opleidingsniveau op de prestatie op de Stroop Color Word Test (SCWT) nagegaan, en werd de normale spreiding van de testcores bepaald. De SCWT meet executief functioneren en daaraan gerelateerde concepten zoals aandachtsfuncties. De statistische analyses waren gebaseerd op een MAAS steekproef van 1,856 gezonde proefpersonen die tussen 24 en 81 jaar oud waren. Multiple regressie analyse toonde aan dat snelheidsafhankelijke SCWT scores in belangrijke mate werden beïnvloed door demografische variabelen, terwijl dit minder het geval was voor maten van nauwkeurigheid (i.c. fouten). De leeftijdsgerelateerde achteruitgang in executief functioneren was meer uitgesproken voor laag opgeleide mensen dan voor hoogopgeleiden, een resultaat dat in de lijn ligt van de ‘cognitive reserve theory’. Traditionele en regressie-gebaseerde normen werden geconstrueerd. Een vergelijking tussen beide methoden suggereerde dat de regressie-gebaseerde methode superieur is aan de traditionele methode, hoewel de traditionele methode de meest gebruikte normatieve procedure is.

Hoofdstuk 3 beschrijft de resultaten van een studie die naging in welke mate leeftijd en leeftijdsextrinsieke variabelen de prestatie op de Verbal Learning Test van Rey (VLT) beïnvloedden. De VLT is één van de meest gebruikte taken om het declaratief geheugen te meten. De leerwoorden werden zowel auditief als visueel aangeboden in MAAS. Samen met de demografische variabelen

werd de invloed van presentatiemodus op testprestatie nagegaan omdat vorig onderzoek suggereerde dat auditief aangeboden materiaal beter wordt onthouden dan visueel aangeboden materiaal, tenminste op korte termijn (het zogenaamde ‘modaliteitseffect’). De VLT scores van een steekproef van 1,855 mensen die tussen 24 en 81 jaar oud waren werden geregresseerd op presentatiemodus, leeftijd, geslacht en opleidingsniveau. De resultaten toonden dat de VLT prestatie reeds afneemt vanaf de jong-volwassenheid, en dat vrouwen en hoogopgeleiden beter presteerden dan mannen en laagopgeleiden. Het effect van presentatiemodus was niet eenduidig. Auditieve presentatie was geassocieerd met een betere prestatie op de VLT trials die steunen op het korte-termijn geheugen, terwijl visuele aanbieding resulteerde in een betere prestatie bij trials die steunen op het lange-termijn geheugen of trials die steunen op een mix van korte en lange-termijn geheugen. Regressie-gebaseerde VLT normen werden geconstrueerd, met correctie voor de relevante onafhankelijke variabelen.

Het doel van Hoofdstuk 4 was te bepalen of leeftijd, geslacht en opleidingsniveau de prestatie op semantische en fonetische Verbal Fluency Tests (VFTs) beïnvloedden. Verschillende VFTs (categorieën Beroepen, Dieren, Letter M) werden afgenomen bij een steekproef van 1,856 mensen. Regressie analyse suggereerde dat leeftijd de prestatie op alle VFTs beïnvloedde, maar dat het leeftijdseffect niet eenduidig was. Voor de Beroepen en Letter M VFTs bleef de testprestatie stabiel tot ongeveer 50 jaar, waarna deze snel achteruitging. Voor de Dier VFT daarentegen verminderde de testprestatie reeds vanaf de jonge volwassenheid. Opleiding had een sterk effect op de prestatie op alle VFTs. Geslachtseffecten werden enkel gevonden voor de Beroepen VFT, waarbij mannen beter presteerden dan vrouwen.

In Hoofdstuk 5 werd de Concept Shifting Test (CST) beschreven. De CST is een aanpassing van de Trail Making Test (TMT). De CST werd ontwikkeld om bepaalde problemen die optreden met de TMT te vermijden. Ten gevolge van bepaalde eigenaardigheden in de lay-out van de TMT is de specificiteit van de maat voor executief functioneren die kan worden afgeleid van deze test niet optimaal. De CST werd afgenomen bij een steekproef van 1,856 personen die tussen 24 en 81 jaar oud waren. Op basis van deze steekproef werd bepaald in welke mate de CST prestatie werd beïnvloed door leeftijd, geslacht, opleidingsniveau en handvoorkeur. Zowel multiple lineaire als (binaire) logistische regressie analyse werd gebruikt om de gegevens te analyseren. De CST prestatie was beter voor vrouwen, daalde als functie van leeftijd, en steeg als functie van opleidingsniveau. Normatieve data die rekening houden met de invloed van deze variabelen werden geconstrueerd met een regressie-gebaseerde normatieve procedure.

In Hoofdstuk 6 werd een papier-en-potlood versie van Sternberg’s Memory Scanning Test (P&P MST) beschreven. Het oorspronkelijke Sternberg paradigma is moeilijk te gebruiken in een klinische omgeving, ondermeer omdat de afname van de originele taak veel tijd kost en de taak te moeilijk is voor ouderen en mensen met cognitieve stoornissen (bvb. dementie). De P&P MST heeft

deze nadelen niet, en is daarom erg geschikt voor gebruik in de klinische praktijk. De P&P MST wordt gebruikt om de snelheid van geheugenscanning en niet-geheugen gerelateerde processen in het werkgeheugen te meten. De P&P MST scores van een steekproef van 1,839 mensen werden geregresseerd op leeftijd, geslacht en opleiding. De resultaten suggereerden dat de snelheid van alle componenten van informatieverwerking in het werkgeheugen van ouderen en laagopgeleiden lager lag dan bij jongeren en hoogopgeleiden. Jongere vrouwen (jonger dan 55) scanden het werkgeheugen sneller dan mannen, maar deze relatie was omgekeerd voor ouderen (ouder dan 55). Normatieve data werden geconstrueerd voor de maten die de snelheid van geheugenscanning en niet-geheugen gerelateerde processen in het werkgeheugen kwantificeren, gestratificeerd naar de relevante onafhankelijke variabelen.

In Hoofdstuk 7 werd de invloed van demografische variabelen op de prestatie op de schriftelijke en mondelinge versies van de Letter Digit Substitution Test (LDST) geëvolueerd. Deze taak is procedureel identiek aan de Digit Symbol Substitution Test. De LDST meet algemene snelheid van informatieverwerking. De schriftelijke en mondelinge versies van de LDST werden afgenomen bij een steekproef van 1,858 volwassenen die tussen 24 en 81 jaar oud waren. Ouderen, mannen en laagopgeleiden presteerden slechter op zowel de schriftelijke als de mondelinge versies van de LDST dan jongeren, vrouwen en hoogopgeleiden. Regressie-gebaseerde normatieve data werden geconstrueerd voor de schriftelijke en mondelinge versies van de LDST.

In Hoofdstuk 8 werd nagegaan in welke mate demografische variabelen de verandering in prestatie na drie jaar in de visuele versie van de VLT, de SCWT en de LDST beïnvloeden. Een grote steekproef mensen die bij baseline tussen 49 en 81 jaar oud waren werd na drie jaar opnieuw getest. De veranderingsscores werden geregresseerd op leeftijd, geslacht en opleidingsniveau. Significante effecten van leeftijd en geslacht werden gevonden voor de SCWT Interferentie, VLT totale herinnering 1-3, VLT uitgestelde herinnering en LDST scores, maar niet voor de VLT onmiddellijke herinneringsscore. De resultaten suggereerden dat de leeftijdsgelateerde achteruitgang in executief functioneren na drie jaar meer uitgesproken was voor mannen in vergelijking met vrouwen. De normale variabiliteit in de verandering in de test scores werd vastgesteld, met correctie voor de relevante onafhankelijke variabelen.

Hoofdstuk 9 geeft enkele concluderende beschouwingen over 1) cognitieve reserve en leeftijd, 2) regressie-gebaseerde versus traditionele normatieve procedures, 3) geschikte onafhankelijke variabelen in normatief onderzoek, en 4) ecologische validiteit en neuropsychologische assessment. Tot slot worden enkele suggesties gegeven om de methoden en aanpak van neuropsychologische assessment in de toekomst te verbeteren. Op basis van de resultaten die werden gepresenteerd in dit proefschrift, kan worden geconcludeerd dat leeftijd een negatieve invloed heeft op de meeste, zometer alle, 'fluid' cognitieve functies. Bij de meeste cognitieve maten werden zowel lineaire als kwadratische leeftijdseffecten geobserveerd. Er werd ook aangetoond dat variabelen zoals geslacht en

opleiding dienen worden in rekening gebracht bij de evaluatie van inter-individuele verschillen in cognitieve functies. De invloed van deze demografische variabelen verschilde voor de diverse cognitieve domeinen. Formele modellering met multiple lineaire regressie analyse (of logistische regressie analyse) voorziet in een elegante aanpak om de relevantie en relatieve bijdrage van meerdere onafhankelijke variabelen op inter-individuele verschillen in cognitie en verandering in cognitieve functies over de tijd te bepalen. Met betrekking tot het bepalen van de normale range van prestatie op neuropsychologische taken zou de traditionele aanpak moeten worden opgegeven en vervangen worden door de regressie-gebaseerde aanpak. In tegenstelling tot de traditionele normeringsaanpak heeft de regressie-gebaseerde aanpak als voordeel (i) dat de populatieparameters accurater kunnen worden geschat, (ii) dat het probleem van niet-gebalanceerde data in de steekproef wordt vermeden, (iii) dat zowel continue, categoriale en dummy variabelen kunnen dienen als predictor van test prestatie, en (iv) dat een algemene aanpak wordt gegeven voor het construeren van normen die van toepassing is in vele situaties. De regressie-gebaseerde normen die werden geconstrueerd in dit proefschrift leveren een belangrijk werkinstrument voor vele neuropsychologen in Nederland en daarbuiten.

Curriculum vitae

Wim Van der Elst werd op 23 oktober 1978 geboren te Aalst (België). In 1996 behaalde hij zijn ASO diploma aan het Instituut voor Katholiek Secundair Onderwijs te Denderleeuw. Hij studeerde in 2001 cum laude af in de psychologie aan de Katholieke Universiteit Leuven. In november 2002 startte hij met een AIO project aan het departement Psychiatrie en Neuropsychologie van de Universiteit Maastricht waarvan de resultaten in dit proefschrift staan beschreven. In 2003 begon hij met de avondopleiding Filosofie (Filosofische academie) aan het Hoger Instituut voor Wijsbegeerte van de Katholieke Universiteit Leuven.

Publications

- Van der Elst, W.**, Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2005). Rey's Verbal Learning Test: Normative data for 1,855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society*, 11, 290-302.
- Van der Elst, W.**, Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006a). Normative data for the Animal, Profession and Letter M naming Verbal Fluency Tests for Dutch speaking participants and the effects of age, education, and sex. *Journal of the International Neuropsychological Society*, 12, 80-89.
- Van der Elst, W.**, Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006b). The Letter Digit Substitution Test: Normative data for 1,858 healthy participants aged 24-81 from the Maastricht Aging Study (MAAS): influence of age, education, and sex. *Journal of Clinical and Experimental Neuropsychology*, 28, 998-1009.
- Van der Elst, W.**, Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006c). The Stroop Color-Word Test: influence of age, sex, and education; and normative data for a large sample across the adult age range. *Assessment*, 13, 62-79.
- Van der Elst, W.**, Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (in press). The Concept Shifting Test: Adult normative data. *Psychological Assessment*.
- Van der Elst, W.**, Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. Assessment of information processing in working memory in applied settings: the paper & pencil memory scanning test. *Submitted to Psychological Medicine*.
- Van der Elst, W.**, Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. Assessment of change in performance on the Stroop Color-Word Test, Verbal Learning Test of Rey, and Letter Digit Substitution Test after a test-retest interval of three years. *Submitted to Psychological Assessment*.