



UiT The Arctic University of Norway

Faculty of Health Sciences

Department of Psychology

Dementia Disease Initiation:

Demographically adjusted norms based on Scandinavian samples and comparison with published norms from North America

Johan Jacob Espenes

A dissertation for the degree of Philosophiae Doctor, July 2023

Table of Contents

Acknowledgements	4
List of Papers.....	5
List of abbreviations.....	6
Abstract	7
1. Introduction	8
1.1 Norms.....	8
1.2 On the culture specificity and heterogeneity of norms	10
1.3 Evolution of norms over time	13
1.4 Traditional and contemporary methods for developing norms	14
1.5 Neuropsychological measures and published norms	17
1.5.1 Trail Making Test (TMT).....	17
1.5.2 Norms on TMT.....	19
1.5.3 Norms on TMT by Heaton (2004) and Tombaugh (2004).....	20
1.5.4 Rey Auditory Verbal Learning Test (RAVLT).....	22
1.5.5 Norms on RAVLT.....	23
1.5.6 Norms on RAVLT by Stricker et al. (2021).....	24
1.5.7 Color-Word Interference Test (CWIT) from the Delis-Kaplan Executive Function System (D-KEFS).....	25
1.5.8 Norms on D-KEFS CWIT and Stroop tests.....	27
1.5.9 Age-adjusted norms on CWIT by Delis et al. (2001).....	27
2. Summary of Introduction	28
3. Objectives.....	28
4. Methods and materials.....	29
4.1 Background on the cohorts	29
4.2 Study samples, joint inclusion and exclusion criteria, and recruitment methods.....	30
4.2.1 Independent Comparison Group for Paper 2	31
4.2.2 Test-retest samples in Paper 2 & 3.....	32
4.3 Materials	32
4.4 General procedures.....	34
4.5 Statistical Analyses.....	35
4.5.1 Comparisons of mean scores	35
4.5.2 Normalization procedures Paper 1 and 3.....	35
4.5.3 Multiple regression analysis Papers 1-3	36
4.5.4 Multivariate regression analysis Paper 2.....	39
4.5.5 Calculating normative performance	40
4.5.6 Assessment of published norms in Scandinavian samples.....	40
4.5.7 Percentiles on skewed measures.....	41
4.5.8 Test-retest reliability.....	42
4.6 Ethics.....	42

5. Summary of results.....	43
5.1 Summary results Paper 1.....	43
5.2 Summary results Paper 2.....	45
5.3 Summary results Paper 3.....	49
6. Discussion	51
6.1 Summary of findings.....	51
6.2 Pattern of age effects in the Scandinavian samples in Papers 1-3.....	52
6.2.1 Do published norms from the US underestimate older participants from Scandinavia?	53
6.2.2 Why might age-effects be lower in the Scandinavian samples?	54
6.3 Effect of education in the Scandinavian samples in Papers 1-3 and comparison with North American norms.....	56
6.3.1 Are Scandinavian norms characterized by lower influence of education?.....	58
6.4 Sex differences in Scandinavian samples in Paper 2	60
6.4.1 Adjustment of sex-differences using Stricker et al. (2021) norms	60
6.5 Clinical implications and suitability of the North American norms in Scandinavian samples ...	62
6.6 Improvements on traditional derived measures by employing regression-based approaches.....	64
6.7 Test-retest reliability in Paper 2 and Paper 3	66
6.8 Methodical considerations and study limitations in Papers 1-3.....	67
6.9 Future directions	70
7. Conclusions	71
8. References	72
9. Papers I-III.....	88

List of Tables and Figures

Table 1 Demographical variables and descriptive statistics for samples in Papers 1-3.....	32
Table 2 Primary and derived measures on TMT, RAVLT and CWIT.....	34
Figure 1 Linear plots of TMT T-scores computed with norms from Heaton et al. (2004), Tombaugh (2004), unadjusted scores and local norms.....	45
Figure 2 Linear plots of RAVLT Trial 7 T-scores computed with Stricker et al. (2021) norms, unadjusted scores, and local norms	48
Figure 3 Plots of T-scores on CWIT-1 and CWIT-3 calculated applying norms from Delis et al. (2001), local norms, and T-scores unadjusted for demographic variables.....	50
Figure 4 Percentage of participants in the Norwegian sample (n = 1011) with a score 1.5 SD below the normative mean (T-score < 35) on CWIT 1-4.....	51

Acknowledgements

I started in the DDI project after my sister Ragna Espenes asked if I would consider working as a research assistant in an interesting research group to gain some more experience in conducting neuropsychological assessments. At the time, I could not have imagined I would apply for the position of Ph.D. candidate in the same group and some years later write the last sentences of my own thesis. For that, I simply lacked the confidence, ambition, and motivation. Luckily, my wonderful supervisor Bjørn-Eivind Kirsebom, who was working towards his own Ph.D. at the time, had enough to spare! Due to his unmatched enthusiasm, inclusive personality, and incredible ability to inspire I reluctantly agreed to apply for the position as Ph.D. candidate at the Department of Psychology. Indisputably, none of this would have been possible without Bjørn-Eivind or Ragna, and I am sincerely grateful for everything.

I also thank my co-supervisor Prof. Knut Waterloo for offering his impressive clinical perspective and keeping everything running behind the scenes at DDI Tromsø. My Ph.D. project was funded by the Department of Psychology at the Arctic University of Tromsø, and I express my gratitude for the privileged opportunity of working towards this degree. I would also like to thank all my colleagues in the national DDI project, the Department of Psychology, and collaborators in Oslo MCI, Gothenburg MCI, and LCBC. In particular, I wish to thank Ingvild Vøllo Eliassen, Erik Hessen, Tormod Fladby, and my colleagues in Tromsø, Ingrid Lorentzen, Stephanie Knudtzon and Grit Richter. Furthermore, my sincerest appreciation goes out to all participants and patients that contributed to this research, and the clinicians at the Department of Neurology, University Hospital of North Norway.

Lastly, I wish to acknowledge all my good friends and family in Tromsø and Tønsberg who supported me. I wish to convey my appreciation to my friends in the Tromsø climbing scene for countless hours of joyous distractions. Most importantly, I thank my loving girlfriend Embla for her patience and perspective and for always believing in me.

List of Papers

Paper 1: Espenes, J., Hessen, E., Eliassen, I. V., Waterloo, K., Eckerström, M., Sando, S. B., ... & Kirsebom, B. E. (2020). **Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years.** *The Clinical Neuropsychologist*, 34(sup1), 110-126.

Paper 2: Espenes, J., Eliassen, I. V., Öhman, F., Hessen, E., Waterloo, K., Eckerström, M., ... & Kirsebom, B. E. (2022). **Regression-based normative data for the Rey Auditory Verbal Learning Test in Norwegian and Swedish adults aged 49–79 and comparison with published norms.** *The Clinical Neuropsychologist*, 1-25.

Paper 3: Espenes, J., Lorentzen I. M., Eliassen, I. V., Hessen, E., Waterloo, K., ... & Kirsebom, B. E. (In review). **Regression-based normative data for the D-KEFS Color-Word Interference Test in Norwegian adults ages 20 to 85.** *The Clinical Neuropsychologist*.

List of abbreviations

MCI	Mild cognitive Impairment
<i>SD, M, SE</i>	Standard Deviation, Mean, Standard Error
TMT	Trail Making Test
CVLT	California Verbal Learning Test
COWAT FAS	Controlled Oral Word Association Test
RAVLT	Rey Auditory Verbal Learning Test
GAMLSS	Generalized Additive Models for Location Shape and Scale
CWIT	Color-Word Interference Test
D-KEFS	Delis-Kaplan Executive Functions System
HRNB	Halstead-Reitan Neuropsychological Battery
GDS	Geriatric Depression Scale
CDR	Clinical Dementia Rating
DDI	Dementia Disease Initiation
LCBC	Center for Lifespan Changes in Brain and Cognition
SCD	Subjective Cognitive Decline
CERAD	Consortium to Establish a Registry for Alzheimer's Disease
ADHD	Attention-Deficit/Hyperactivity Disorder
US	The United States

Abstract

Background: Norms are necessary for interpreting neuropsychological test scores. In Norway and Sweden, there is a lack of local norms on neuropsychological tests. Consequently, Scandinavians are frequently norm-referenced to participants from North America tested decades prior. However, factors such as time of assessment, language, population differences in average performance on cognitive tests, and cultural differences in education and health related factors are sources of heterogeneity in norms. Thus, there is a need for neuropsychological test norms based on Scandinavian populations. In this thesis, the primary aim was to develop norms on the Trail Making Test (TMT), Rey Auditory Verbal Learning Test (RAVLT) and Delis-Kaplan Executive Functions System Color-Word Interference Test (D-KEFS CWIT) based on Scandinavian samples of healthy adults. The secondary aim was to assess frequently used norms from North America in Scandinavian samples.

Methods: Based on healthy adult participants from Norway and Sweden, we modelled pertinent effects of age, education, and sex using regression-based norming procedures in Papers 1-3. We provide normative calculators to aid clinicians and researchers in the computation of normed scores. We assessed overall distributions and performed multiple regression analyses with demographical variables as predictors to assess whether North American norms adequately adjusted for demographical variables in *T*-scores. We calculated test-retest reliability indices (Paper 2 & 3) and assessed differences in the estimated rate of impairment using local norms and North American norms (Paper 3).

Results and conclusions: In the Scandinavian samples, age, education, and sex was significantly related to scores on TMT, RAVLT and D-KEFS CWIT. Results indicated adequate reliability for the most prominent subtests. Compared to previous studies, education appeared to explain less variance in scores in the Scandinavian samples. Compared to previously published norms from North America, results indicated less difference between young and old participants, participants with high and low educational attainment, and less difference between men and women on RAVLT. These discrepancies may be in part due to cultural differences in education and health-related factors. In future research, there is a need for harmonized representative samples to assess whether these results represent generalizable differences characterizing healthy Scandinavians. The North American norms were too lenient for elderly individuals and/or those with low educational attainment, and too strict for individuals with high educational attainment and/or younger individuals. Results from Paper 3 suggest that this may have clinical implications for the accurate assessment of cognitive functions in Scandinavian samples.

1. Introduction

In the coming years, most high-income countries face a demographic change characterized by a rapidly growing proportion of elderly (World Health Organization, 2018). While this is celebrated as an accomplishment of modern society, it is unfortunately also associated with an increased prevalence of dementia and mild cognitive impairment (MCI). A recent study estimated that over 101 000 individuals in Norway suffer from dementia which corresponds to about 14.6% of all individuals over the age of 70 (Gjora et al., 2021). However, by 2050 the prevalence is expected to more than double. Furthermore, an overwhelming 35.5% of people over the age of 70 in Norway were estimated to have cognitive impairments representative of MCI. The necessary health care for patients with dementia is associated with huge financial costs which were estimated to a total of \$1.3 trillion in 2018 and was recently named the costliest disease for Norwegian health care (Kinge et al., 2023; Wimo et al., 2023). Cognitive deficits are a known risk factor for further progression to dementia (Espinosa et al., 2013; Michaud, Su, Siahpush, & Murman, 2017). Thus, research on dementia, in addition to accurate assessment of cognitive deficits, is of great significance for societies in Norway, and may be important for an increasing number of individuals in the future. For this, we need tests that are reliable and valid for accurately assessing cognitive decline in adults and elderly (Bondi et al., 2014).

1.1 Norms

To aid neuropsychologists and other users of neuropsychological tests we use norms to interpret scores. Norms are points of reference for what is considered the average performance in a defined population (Strauss, Sherman, & Spreen, 2006). The defined population often consists of healthy participants without apparent pathologies. Thus, great deviation from the average performance in this population informs the clinician about the relative performance of the patient, which the clinician may further interpret as likely associated with neuropathology or other disorders. For instance, knowing that a patient remembered 7 out of 15 words is not informative until we also know the norm. The norms may, for instance, convey that the expected score is 9 and that people on average deviate from this by 3 words. With this information, the clinician can estimate that about 25% of people remember 7 or fewer words, and thereby correctly interpret the patient's score as a slightly below average result.

However, commonly used neuropsychological tests are known to vary with age, sex, and education (Mitrushina, Boone, Razani, & D'Elia, 2005; Strauss et al., 2006). That is, scores on

a test may, for instance, on average be lower for older individuals in the healthy population. As a result, the range describing normal scores for a 75-year-old woman with 12 years of formal education is not the same as for a 50-year-old man with 19 years of formal education. Therefore, to get an accurate assessment of whether someone performed within the expected range on a given test, the results must be analyzed considering the expected performance given the specific demographical background of this individual. In other words, the norms need to be demographically adjusted.

When performance on a cognitive test is compared to a defined population it is formally referred to as a norm-referenced interpretation or norm-referenced test (Arne Evers, 2012). Informally, and in most cases, this is simply referred to as norms. In contrast, clinicians may want to investigate what raw score on a neuropsychological test best separates healthy participants from some sample with known pathology, or what raw scores on average are associated with, for instance, impaired driving ability. This is referred to as diagnostic norms or criterion-referenced tests.

In Norway, most neuropsychological tests in use by clinicians and researchers have available norms. However, these norms are often based on foreign participants sometimes tested several decades ago. Ryder (2021) reviewed the use of neuropsychological tests in Norway and found that on the most popular tests in use by Norwegian psychologists, local norms and information on validity and reliability for these measures were lacking. In practice, this means that most administrators of neuropsychological tests in Norway are assessing patients and norm-referencing them to participants from other countries without knowledge on the validity of these measures or norms in a local setting. This convention entails, for instance, that a participant from Norway who is 70 years old and has a master's degree is compared to 70-year-olds from the US with a master's degrees.

This practice is not optimal because there may be fundamental differences between participants from Norway compared to other countries. As will be discussed in more detail in the next sections, these differences could manifest due to systematic differences in the average performance between countries on cognitive tests, the magnitude of the association between demographic variables and test scores may differ between countries, cohort differences arise due to the time of assessment, and cultural factors influence the suitability of test material and attitudes towards testing. Furthermore, method biases due to incompatible samples preclude representative norm-referencing, and sub-optimal methods for developing norms may bias the normative estimates. This is important considering neuropsychological

test scores are frequently used to inform clinicians in important decisions including but not limited to future treatment for the patient, eligibility for driving, insurance, and eligibility for social benefits. Thus, research is needed on the implications on using imported norms in a local setting, and the development of local norms is a priority.

1.2 On the culture specificity and heterogeneity of norms

A few studies have investigated the consequences of applying published international norms on neuropsychological measures in Scandinavian countries. These studies motivate the continued development and usage of local norms in Scandinavia. First, Fernandez and Marcopulos (2008) compared published norms on the Trail Making Test (TMT) from several western countries and found evidence of considerable heterogeneity between norms. For instance, when comparing the normative mean described in norms from the US and in norms from Sweden, the difference in age-matched groups varied between 0.8 standard deviations (*SD*) and 1.4 *SD* across all considered age groups. The US norms were consistently stricter. Consequently, applying the US norms in the Swedish sample would result in misclassification of cognitive test performance, in which a substantial part of the Swedish normative sample were mistakenly classified as impaired.

Raudeberg, L. Iverson, and Hammar (2019) compared the use of Scandinavian norms and US norms for diagnosing executive dysfunction in a clinical sample of Norwegian patients with Schizophrenia using the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS). RBANS is a brief test covering several cognitive functions including immediate memory, attention, language, visual cognition, and delayed memory. Results indicated that Norwegian patients scored significantly better using US norms compared to Scandinavian norms on all summary measures except attention. Overall, this had profound implications for the accurate assessment of cognitive dysfunctions in this group as results indicated that 20% fewer patients were classified as impaired applying the US norms.

In a sample of cognitively healthy men and women from Norway, Egeland et al. (2005) assessed whether the mean scores obtained by this group on the California Verbal Learning Test (CVLT) approximated the normative mean reported in the original published norms from the US. Results indicated that women performed significantly worse compared to the original norms on the most prominent measures from the CVLT. On average, the Norwegian women obtained scores 1.1 *SD* below the normative mean on delayed cued recall, and 0.7 *SD* below the normative mean on delayed free recall. For men however, the original US norms adequately reflected the mean scores obtained by this group on the CVLT.

Recently, we investigated differences between our newly developed local norms, previously published local norms, and previously published foreign norms from the US in a sample of cognitively healthy adults from Norway and Sweden on the Controlled Oral Word Test (COWAT FAS) (Lorentzen et al., 2023). Our results indicated that the published norms from the US did not adequately adjust for the effects of age or education in scores in the Scandinavian sample. As a result, the older participants with high educational attainment exceeded the normative expectations set in the US norms and on average these participants received too high scores relative to the expected normative mean. The differential effect of education on COWAT FAS scores could be due to differing frequency of words starting with F, A and S between English and Norwegian causing an in-compatibility of test demands (Lorentzen et al., 2023).

In summary, these studies suggest that local norms in Norwegian and Swedish samples may be needed and highlight some of the implications of using foreign norms in a local setting. Similar results have been reported in international studies outside Scandinavia as well. Hestad et al. (2016) assessed how well US norms based on a sample of healthy African Americans could correct for demographical variables and assess cognitive functioning in Zambia using the tests from the Halstead-Reitan Neuropsychological Battery (HRNB) (Heaton, 2004). The norms adjusted for much of the effects of education on scores, but contrary to the US, men performed better than women on measures of verbal episodic memory in Zambia. Furthermore, applying a criterion of 1 *SD* below the normative mean as an indicator of impairment, as much as 68% of the Zambian sample fulfilled criterion for impairment, compared to the expected base rate of 16%. Thus, the percentage of Zambian participants estimated to be 'impaired' using US norms were highly inflated compared to the expected base rate.

Furthermore, research on differences in cognitive test performance between countries and regions imply that the exchange of norms may not be valid due to consistent population differences in the performance on cognitive tests (Skirbekk, Loichinger, & Weber, 2012; Weber, Skirbekk, Freund, & Herlitz, 2014). Differences in cognitive test performance and corresponding impairment rates are reported in adults and elderly between European countries in large scale studies on diverse cognitive outcomes (Barbosa, Midão, Almada, & Costa, 2021; Formanek, Kagstrom, Winkler, & Cermakova, 2019). In these large scale studies, which were harmonized to reduce incomparability due to method biases, participants from Scandinavian countries displayed the best performances on tests and lowest impairment rates

compared to other European countries, even after adjusting for sociodemographic variables (Formanek et al., 2019).

The reasons for differences in neuropsychological performance between countries, regions, and cultures, include so-called ‘life-course differences’ that affect the performance on cognitive tests at a population-level (Skirbekk et al., 2012). These include factors related to education such as educational length and quality, nutrition across the lifespan, physical and social activities, pollution and adverse health conditions, socioeconomic status, and cognitive stimulation at work (Lövdén, Fratiglioni, Glymour, Lindenberger, & Tucker-Drob, 2020). Furthermore, differences between countries may manifest due to cultural factors influencing the test strategies employed by individuals in these countries. Participants and patients may vary in their ‘test-wiseness’, referring to the shared expectations that test situations entail working fast, silently, and accurately (Nell, 1999). For instance, studies have reported that participants from Russia were more likely to prioritize accuracy over speed on timed cognitive tests compared to participants from the US (Hayden et al., 2014). Likewise, Ojeda, Aretouli, Peña, and Schretlen (2016) report that participants from Spain over the age of 40 also prioritized accuracy over speed, which they correlate with the cultural environment surrounding the fascist regime of Franco which putatively heavily penalized making errors in Spanish schools. We are not aware of any similar studies conducted in Scandinavian countries; however, this indicates that differing strategies and attitudes towards testing may differ between countries, and within countries for different age cohorts.

As a result of differences between countries, demographical variables commonly adjusted for in norms may have differential effects on scores. That is, the specific impact or magnitude age, education, and sex have on scores may vary between countries. For instance, the difference between young and old participants may be greater in some countries compared to others due to accessibility and quality of health care, work environment, and socioeconomic status which affects the average observed rate of cognitive decline (Lövdén et al., 2020). Internationally, Hayden et al. (2014) reported that the detrimental effect of age on verbal episodic memory scores were stronger in Russia compared to the US, even after adjusting for variables known to impact scores such as cardiovascular disease (CVD), depression, and education. Furthermore, Rivera et al. (2015) reported differing effects of education between Spanish speaking countries in Latin America that were harmonized to follow the same procedures for recruitment, sampling, inclusion and exclusion criteria. Adjusted for any differences in age, higher education was positively related to scores on a Stroop task in

Argentina, Chile, Mexico, and Paraguay, but not in Puerto Rico, Peru, or El Salvador. In other words, even in countries that share the same language and other cultural factors (Rivera et al., 2015), using the same procedures, and attempting to recruit similar individuals, there are differing effects of education on scores. Lastly, in a meta-analysis Asperholm, Nagar, Dekhtyar, and Herlitz (2019) found that sex-differences on episodic memory tests greatly differed between countries and was correlated with social progress indicators such as gender equality, gross domestic product per capita (GDP), and educational attainment. Consequently, even though norms may be adjusted for age, education, and sex, this does not mean that norms can be transferred to other countries with equivalent validity.

1.3 Evolution of norms over time

Irrespective of the culture specificity of norms, it is well known that the time of assessment is important as population-level performance within a country on cognitive tests are known to change over time. In the past century it has been observed that adults and elderly tend to perform better than participants of the same age tested in the preceding decades. This secular increase in cognitive test performance from generation to generation is referred to as the Flynn effect (Flynn, 1987). While the exact reason for the Flynn effect is still debated, the Flynn effect is generally thought to be due to population-level improvements in education, nutrition and health care, economic conditions, decline in mortality and reduced family size (Skirbekk et al., 2012; Williams, 2013). The Flynn effect has been observed for various cognitive abilities including tests of executive functions (Dickinson & Hiscock, 2011) and verbal episodic memory (Baxendale, 2010; Weber et al., 2014), but it has mostly been studied based on standardized measures of general cognitive ability like Ravens Progressive Matrixes (Williams, 2013). Moreover, the gain in cognitive ability from generation to generation is known to vary in magnitude between countries (Hessel, Kinge, Skirbekk, & Staudinger, 2018). In fact, in countries like Norway some studies show a stagnation or even reversal of the Flynn-Effect (Bratsberg & Rogeberg, 2018; Sundet, Barlaug, & Tojussen, 2004; Teasdale & Owen, 2005). This may further limit the generalizability of normative estimates between countries. In recognition of the secular gains in cognitive performance, the available guidelines for test developers and users indicate that norms older than 15 years are considered potentially out-of-date and warrant caution, while norms older than 20 years are considered inadequate (Arne Evers, 2012).

1.4 Traditional and contemporary methods for developing norms

In traditional norms, norm developers use demographical variables like age or sex to define discrete subgroups and estimate the norm statistics such as mean (M), SD , or percentiles directly from each separate subgroup (Aarts & Oosterhuis). For instance, if scores on a given neuropsychological measure are known to differ between men and women, the normative sample is split according to sex and the norm statistics are estimated from the two separate subgroups. However, for continuous variables such as age there is no way to reliably estimate the test-score distributions of subgroups representing every discrete value of age in the sample without requiring enormous sample sizes. Instead, norm developers are forced to stratify the sample in discrete age bins. For instance, norm developers may want to stratify participants in ten-year intervals, e.g., participants between 50-59 years, 60-69 years and so on. Because of this, traditional norms are sometimes referred to as discrete norms or stratified norms.

Compared to more modern ways of developing norms, traditional norms require large sample sizes to obtain adequate precision. This is because the distribution of scores in each separate subgroup is used to calculate the norm statistics. As a result, each subgroup needs to be representative of the population from which it is drawn (e.g., women between age 50 and 59 years). Given a lenient sample size requirement of 40 participants for approximating the normal distribution according to the central limit theorem (Mitrushina et al., 2005), splitting by sex and 8 different age bins means that 16 subgroups with approximate normal distributions need to be estimated. As a result, the minimum sample size needed to estimate the norms is 640. As expected, if scores are known to vary according to education as well, the sample size requirement multiplies further.

Another critique of traditional norms is that moving from one normed subsample to the next can have large implications for the interpretation of individual scores (Parmenter, Testa, Schretlen, Weinstock-Guttman, & Benedict, 2010). For instance, an individual that is 59 years old might be norm-referenced to the age 50-59 years subgroup on one assessment but turns 60 years before the next and is therefore compared to individuals aged 60–69 years. This might drastically change the interpretation of this individual's score. This is known as the 'edge of cohort effect' (Crompvoets, Keuning, & Emons, 2021). Therefore, traditional norms are either imprecise due to the crude stratification of demographic variables, or resource demanding due to the large sample size requirements. As a result, traditional norms are prone to misleading results (Van Breukelen & Vlaeyen, 2005).

More modern extensions of the traditional method include the overlapping interval strategy (García-Herranz et al., 2022) sometimes referred to as interval superposition (Llinàs-Reglà et al., 2017; Specka et al., 2022) or over-lapping cell procedure (Engedal et al., 2023; Pauker, 1988). Like traditional norms, these norms present raw scores corresponding to commonly used percentiles (50th, 16th, 7th, 2nd) for discrete age-ranges (for instance, age 60-62 years) further stratified by sex or educational level if needed. However, unlike the norms already discussed, the percentiles are tabulated based on scores for participants +/- 5 years surrounding the age interval (i.e., from 56 up until 66 years of age). As the name suggests, this produces over-lapping cells which increases the number of age-intervals in the norms. Recently, such norms were computed for the Norwegian version of the Mini Mental State Examination (MMSE) (Engedal et al., 2023).

As an alternative to the traditional approach, Zachary and Gorsuch (1985) proposed the use of linear regression analysis for estimating norms. In regression-based norming, test scores on neuropsychological measures are regressed on relevant variables for test performance like age and sex. The resulting regression coefficients describe the average linear association between the predictors and the neuropsychological measure based on the entire sample. To calculate regression-based norms we compare the obtained score by the participant with the predicted score estimated from the regression coefficients. The residual (i.e., the difference between the obtained score by the individual and the predicted score) is then standardized by dividing the residual with the *SD* of the residual (i.e., average departure across the sample). Assuming the standardized residuals follow a normal distribution, scores can then be expressed and readily interpreted in any standardized and normalized format such as *T*-scores, *Z*-scores, or percentiles.

The main advantages of regression-based norming are that the entire sample is used to calculate the normative statistics. This is much more efficient requiring up to 5.5 times smaller sample sizes compared to traditional norms for equal precision (H. E. Oosterhuis, van der Ark, & Sijtsma, 2016). Regression-based norms are also less influenced by unbalanced samples and predicted scores may be estimated for all combinations of demographic variables, even those which were not present in the normative sample (Kleinbaum, Kupper, Nizam, & Rosenberg, 2013; Wim Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2006). Also, because linear regression analysis is well suited for continuous variables, there is no need for the arbitrary stratification of predictors, resulting in a smooth estimation of the normed scores at all possible values of the predictors. Crucially, this solves the edge of cohort

effect in norms. However, for estimates from regression-based norming to be valid norm developers need to take great care that the assumptions of linear regression analysis are fulfilled, or estimates will be biased (Crompvoets et al., 2021; H. E. Oosterhuis et al., 2016).

Simple linear regression analysis can be expanded to instances where the assumptions of linear regression do not hold. That is, in case of non-normally distributed residuals, homoscedasticity, non-linearity, correlated errors or repeated measures (James, Witten, Hastie, & Tibshirani, 2013). For instance, norms are developed for tests with correlated outcomes which are difficult to norm in both traditional and regression-based norms. On the Rey Auditory Verbal Learning Test (RAVLT), the regression-based approach is typically conducted as a series of univariate analyses (Stricker et al., 2021; Testa, Winicki, Pearlson, Gordon, & Schretlen, 2009). However, for multi-trial measures such as the RAVLT, the same participants complete all trials, and the trials are therefore expected to be highly correlated. Thus, Van der Elst, Molenberghs, van Tetering, and Jolles (2017) recommended a multiple multivariate regression-based approach for norming scores from the RAVLT. In this approach, multiple covariates are analyzed on all RAVLT measures jointly. As argued by Van der Elst et al. (2017), there are three advantages associated with this method over the conventional regression-based procedure on the RAVLT; 1) it decreases the likelihood of chance capitalization (i.e. norming scores according to covariates that are not related to performance); 2) it is parsimonious; 3) the correlated nature of trials is accounted for in the analyses, leading to more precise fixed effect (i.e. mean) estimates thus reducing the sample size requirement while attaining comparable precision. Additionally, the multivariate approach allows for formal testing of hypotheses about the evolution of performance over successive trials that is not readily examined with univariate analyses. For instance, if the effect of education varies on successive parts of the RAVLT. Compared to available methods for analyzing longitudinal data like repeated measures ANOVA and MANOVA, the multiple multivariate method is more flexible, allowing for adaption of the estimation method, as well as flexible modelling of the variance-covariance structure to fit the data at hand (W. Van der Elst, Molenberghs, van Tetering, & Jolles, 2017). Compared to the Linear mixed model which allows for fitting fixed and random effects and accounting for missing data, the multiple multivariate approach is preferred on the RAVLT because the fixed effects (means) are of interest to predict scores and there is very little attrition between trials on the RAVLT (W. Van der Elst et al., 2017).

Recently, more advanced methods suitable for non-normally distributed, homoscedastic, and non-linear data has been used for developing norms. These are generalized linear models such Generalized Additive Models for Location Shape and Scale (GAMLSS) and semi-parametric continuous norming approaches accessible through the package ‘cNORM’ in R (Gary, Lenhard, & Lenhard, 2021; Lenhard, Lenhard, Suggate, & Segerer, 2018). However, norming scores with GAMLSS requires expert knowledge in correctly modelling each parameter in the model. Efforts have been made to make such methods accessible for norm developers without the prerequisite knowledge in statistical modelling, but currently the only guidelines that exist are suitable for single continuous predictors (Lieke Voncken, Casper J Albers, & Marieke E Timmerman, 2019). Likewise, semi-parametric continuous norming may be best suited for single continuous predictors due to the great number of polynomial terms estimated (Lenhard et al., 2018). Consequently, regression-based norms based on general linear models remain the most popular method for developing norms in recent years.

1.5 Neuropsychological measures and published norms

So far, I have reviewed general considerations for using norms and considered studies comparing cognitive performance across regions and countries that question the practice of importing or exchanging norms. Secondly, I have reviewed the two main practices for developing norms, namely traditional norms, and regression-based norms.

In this thesis I sought to develop regression-based norms on three popular neuropsychological measures that were identified to have lacking normative data and corresponding validity based on a Scandinavian population (Ryder, 2021). Therefore, in the following sections I will present previously published norms on the measures Trail Making Test (TMT), Rey Auditory Verbal Learning Test (RAVLT), and the Color-Word Interference Test (CWIT) from the Delis-Kaplan Executive Function System (D-KEFS).

1.5.1 Trail Making Test (TMT)

In Paper 1 we analyzed scores on the TMT. First introduced in 1938 as ‘Partington’s Pathways Test’, the TMT has a long history in the field of neuropsychology (Battery, 1944; Partington & Leiter, 1949). TMT remains as one of the most popular tests used by neuropsychologists internationally (Kreutzer, DeLuca, & Caplan, 2011; Rabin, Barr, & Burton, 2005) and in Norway (Vaskinn & Egeland, 2012). Here, we used the TMT version first included in the HRNB for comprehensive psychological assessment (Reitan & Wolfson,

1985). And as far as we are aware, this version is free to use and distribute in clinical and research settings.

Briefly, the TMT is divided into two parts: TMT-A involves connecting numbers and TMT-B involves alternating between connecting numbers and letters. While there is no definite consensus on the cognitive processes underlying TMT performance and the terminology to describe them, TMT-A is considered to primarily measure speed of visual search and perceptual speed (Sánchez-Cubillo et al., 2009). TMT-B is thought to be even more demanding on speed of visual search and perceptual speed due to increased distance and more stimuli distracting participants (Gaudino, Geisler, & Squires, 1995). Additionally, due to the alternating between numbers and letters, TMT-B involves working memory and task-switching (i.e., the ability to fluently switch attention between competing tasks) and is conceptualized as a measure of executive control (Sánchez-Cubillo et al., 2009; Varjadic, Mantini, Demeyere, & Gillebert, 2018). TMT-A and TMT-B has been shown to have acceptable convergent validity and reliability in a Swedish sample of elderly (Pellas & Damberg, 2021).

Derived measures on the TMT are those which are computed based on performance on the basic measures TMT-A and TMT-B. Derived measures on the TMT are primarily the difference score TMT B-A (Lezak, Howieson, Bigler, & Tranel, 2012) and the ratio score TMT B/A (Arbuthnott & Frank, 2000). These were designed to isolate the additional executive demands associated with TMT-B from the lower order perceptual demands associated with TMT-A. Participants usually take longer to complete TMT-B, and normed TMT B-A scores provide information on the average discrepancy between TMT-B and TMT-A adjusted for pertinent demographical variables. An elevated score on TMT B-A is interpreted as difficulties with the additional demands associated with TMT-B (i.e., working memory and task-switching). As such, TMT B-A is conceptualized as a relatively pure measure of task-switching and executive control (Sánchez-Cubillo et al., 2009). TMT B/A provides a ratio between the completion time on TMT-B and TMT-A. The criterion validity of TMT B/A is less clear (Martin, Hoffman, & Donders, 2003), although some report it to be an even purer measure of executive control than TMT B-A due to weaker associations with demographic variables (Arbuthnott & Frank, 2000).

The TMT is recommended as a screening tool for neurological integrity (Reitan & Wolfson, 1994). TMT is suited for this purpose because performance on the TMT draws on diverse neural underpinnings (Varjadic et al., 2018). As a result, many forms of neuropathology may

disturb task performance. This makes TMT a sensitive tool, but not highly specific for any particular type of neuropathology. Measurement of cortical thickness, functional imaging during task completion, and lesion-symptom mapping studies have implicated the involvement of several different brain regions in the prefrontal and parietal cortex, including but not limited to: the left dorsolateral prefrontal cortex; left dorsomedial prefrontal cortex; cingulate sulcus and the intraparietal sulcus (Lee, Wallace, Raznahan, Clasen, & Giedd, 2014; Miskin et al., 2016; Moll, de Oliveira-Souza, Moll, Bramati, & Andreiuolo, 2002). For a review, please see Varjacic et al. (2018). The wide involvement of different brain regions might support the notion that TMT is a sensitive but not highly specific. As such, the TMT has demonstrated utility in many different clinical groups (Bezdicek et al., 2012; Bezdicek et al., 2017; Biundo et al., 2013; Cerezo García, Martín Plasencia, & Aladro Benito, 2015; Martins da Silva et al., 2018; Sparding et al., 2015; Strauss et al., 2006; Wei et al., 2018).

1.5.2 Norms on TMT

To get an overview of the available normative studies on the TMT for the purpose of this thesis I searched Pubmed for [(“Trail Making Test”) AND (normative)] which generated 214 results. By design, this search identified common variations of normative including “norm” and “norms”. Out of the 214 results, 73 were identified to be unique norms based on healthy participants. The oldest norms on the TMT in this selection were published in 1965 and the newest in April 2023. As might be expected, the studies varied according to norming methodology (either traditional or regression-based), detail and breadth of reported statistics, screening procedures and eligibility criteria, sample characteristics, and variables stratified for or adjusted for in the norms.

Notably, only found three normative studies that were based on a Scandinavian sample. Firstly, Nielsen, Knudsen, and Daugbjerg (1989) analyzed results on the TMT with 101 participants between the age 20-54 years from Denmark. The sample consisted of volunteers recruited from the laundry department at Copenhagen University Hospital. This Danish study might be of limited generalizability due to a traditional normative methodology in combination with a small sample size, potential biases associated with the sampling, restrictive age span (20-54 years), no adjustment for education, and that it may be considered old and outdated according to the available guidelines (Arne Evers, 2012).

Secondly, Fällman, Lundgren, Wressle, Marcusson, and Classon (2020) developed traditional norms exclusively for TMT-A based on longitudinal analysis of 207 participants from Sweden. Participants were all born in 1922 and were tested three times at ages 85, 90, or 93

years. Thus, the norms are adjusted for age in three levels (85, 90, 93) and education in three levels (4-9, ≥ 10 , combined). This study is unique in that it presents norms for a very old sample tested three consecutive times. However, the actual applicability of the norms is niche because the study only presents norms on TMT-A for three discrete ages.

Thirdly, Selander, Wressle, and Samuelsson (2020) made traditional norms adjusting for age on the TMT based on a Swedish sample of 410 healthy participants between 20 and 89 years old. They present norms adjusted for age in four levels (20-39 years, 40-59 years, 60-69 years and ≥ 70 years). The disadvantage of these norms is related to the traditional approach, and that they do not adjust for educational attainment which is commonly known to influence scores on the TMT (Stuss, Floden, Alexander, Levine, & Katz, 2001).

Generally, it is reported that younger participants with more education perform better on the TMT (Fábián, Kenyhercz, Bugán, & Andrejkovics, 2023; Lojo-Seoane et al., 2023; Málišová et al., 2022; Rodríguez-Lorenzana et al., 2021; Specka et al., 2022). Linear and quadratic effects of age and/or education have been reported on the TMT (e.g., steepening increase of scores with higher education) (Cavaco et al., 2013; Magnúsdóttir, Haraldsson, & Sigurdsson, 2021; Rodríguez-Lorenzana et al., 2021). Sex differences on the TMT are somewhat unreliable. Most studies do not find any significant sex differences (Mitrushina et al., 2005), however, others have found that women perform worse than men (Cavaco et al., 2013; García-Herranz et al., 2022; Magnúsdóttir et al., 2021) while some find that men perform worse than women (Suzuki et al., 2022). Overall, the effect sizes associated with any significant sex difference is typically lower than those associated with age or education.

Scores on the derived measures TMT B-A and TMT B/A are associated with age and education but the effect sizes are typically smaller than on basic measures of TMT (Arbuthnott & Frank, 2000; Bezdicek et al., 2012; Cavaco et al., 2013; Hester, Kinsella, Ong, & McGregor, 2005; Lojo-Seoane et al., 2023; Specka et al., 2022).

1.5.3 Norms on TMT by Heaton (2004) and Tombaugh (2004)

In Paper 1 we assessed published norms from North America by Heaton (2004) and Tombaugh (2004) in a Norwegian and Swedish sample. Both norms were cited as norms recommended for use in the Norwegian version of the TMT by the Norwegian health authorities (Strobel et al., 2018).

Tombaugh (2004) published norms on the TMT based on a sample of 911 Canadian participants. The normative sample included participants aged 18 to 89 years ($M = 58.5$, $SD =$

21.7), and years of education ranged from 5 to 25 ($M = 12.6$, $SD = 2.6$). It is not stated when assessments were done, however, parts of the normative sample were sourced from a study published in 1994. Approximately half the sample were female. The sample comprised community dwelling volunteers recruited via public spaces and word of mouth. Participants were screened using the Mini Mental State Examination (MMSE) and the Geriatric Depression Scale (GDS) with cut-off scores >23 and <14 , respectively. Participants were excluded if they reported a history of psychiatric illness, head injury, neurological disease, or stroke. Somatic and psychiatric health information was based on self-report.

The TMT was administered according to standardized procedures (Strauss et al., 2006).

The Tombaugh (2004) norms were made with a traditional norming approach. Age was stratified in eleven levels (18-24, 25-34, 35-44, 45-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85-89) split by education in two levels (0-12, >12) exclusively for participants over 60 years of age. As a result, sample sizes were frequently under 40 for some combinations of age and education (lowest $n = 16$, $n = 13$).

In sum, the Tombaugh (2004) likely represented an improvement over contemporary norms because they were more comprehensive (i.e., presented scores for a wide range of age and education) and were based on a larger sample size than previous norms at the time. The norms encompassed the entire adult life span and were adjusted for education. However, compared to modern norms, they are hindered by shortcomings that question the validity of the norms. Namely, lenient exclusion criteria by today's standards; convenience sampling methods; traditional norming approach and low sample size in certain stratifications of age and education.

The Heaton (2004) norms assessed in this thesis were based on a normative sample consisting of 634 Caucasian participants. The normative sample was compiled from multiple studies and were tested during a period of 25 years preceding 2004. Reportedly, most of the sample were tested prior to 1991. All participants received monetary compensation for participation and recruitment likely followed a convenience sampling strategy. The normative sample in Heaton (2004) spans most of the adult life span (20-85) and covers a wide range of educational attainment (0-20). An advertised advantage of the Heaton (2004) norms was the co-norming of the HRNB, however this has been shown to be of limited clinical utility compared to combining norms from multiple studies (Rohling et al., 2015). To normalize measures, Heaton (2004) converted raw scores to scaled scores ($M = 10$, $SD = 3$). Then, norms were estimated using multiple regression analysis with fractional polynomials

adjusting for age, sex, and education on the scaled scores (Heaton, 2004). The norms are then presented in discrete bins for age, education, and sex thereby apparently reducing the resolution of the norms, although there are no studies assessing the practical implications of this approach compared to the continuous approach typically used in regression-based norms (Van Breukelen & Vlaeyen, 2005; Zachary & Gorsuch, 1985). Generally, test procedures on the TMT align with standardized procedures described in Strauss et al. (2006). On the TMT, age reportedly accounted for 25% of the variation of scores on TMT-A and TMT-B, and education accounted for 10% on TMT-A and 16% on TMT-B, unadjusted for age (Heaton, 2004).

1.5.4 Rey Auditory Verbal Learning Test (RAVLT)

In Paper 2 we analyzed scores on the RAVLT which is a classical neuropsychological test for verbal episodic memory (Boake, 2000; Rey, 1958). The test consists of fifteen unrelated nouns that are read aloud to the participant (list A), to which the participant is asked to correctly recall as many words as possible (Trial 1). This procedure is repeated 5 times. Then, a word list with fifteen new words (list B) is read out and again the participant is required to recall as many words as possible from this new list. Immediately after, the participant is asked to recall words from list A once more without any renewed presentation (Trial 6), and finally once more after a thirty-minute timed delay (Trial 7).

The dependent measures on the RAVLT are number of correctly recalled words at each point in the test procedure (i.e., Trials 1 through 5, list B, Trial 6, and Trial 7). The RAVLT is thought to reflect acquisition or learning, attention, working memory, short term memory, retention, retrieval (Ivnik et al., 1992). Furthermore, derived measures are calculated to assess isolated parts of the processes that support performance on verbal episodic memory tests such as learning over time (LOT) or long-term percent retention (LTPR) (Vakil, Greenstein, & Blachstein, 2010).

An advantage associated with the RAVLT is that it is free to use and adapt unlike other popular tests such as the CVLT-II. Furthermore, performance on both tests been shown to be highly correlated in healthy participants (Beier, Hughes, Williams, & Gromisch, 2019; Stallings, Boake, & Sherer, 1995). Compared to other verbal learning tests like the CERAD word list (Consortium to Establish a Registry for Alzheimer's Disease) (Fillenbaum et al., 2008), the RAVLT may offer a more detailed analysis of memory functions and, importantly, the RAVLT is sufficiently hard for most participants. The advantage of a sufficiently challenging test is that ceiling effects which skew test-score distributions occur less frequently

(Kirsebom et al., 2019; Uttl, 2005). On the other hand, this means that the RAVLT might be experienced as too hard by some participants thereby reducing motivation and test effort (Poreh, Tolfo, Krivenko, & Teaford, 2017).

Scores on the RAVLT and other measures of verbal episodic memory correlate with cortical volume in the medial temporal lobe, left lateral temporal lobe, and overall hippocampal integrity (Molinuevo et al., 2011; Saury & Emanuelson, 2017). As such, RAVLT is useful for assessing amnesic deficits due to Alzheimer's Disease (AD) and has shown good diagnostic accuracy for identifying patients with prodromal AD (Vuoksima, McEvoy, Holland, Franz, & Kremen, 2020) and progression to AD dementia (Eckerström et al., 2013). Furthermore, derived measures enable process-oriented interpretation and have specific clinical utility, for instance, interference effects may be sensitive for frontal lobe dysfunctions (Torres, Flashman, O'leary, & Andreasen, 2001; Vakil et al., 2010).

1.5.5 Norms on RAVLT

To get an overview of the available normative data on the RAVLT I searched Pubmed for: [(Rey "Auditory Verbal Learning Test") AND (normative)]. This search returned 73 results out of which 24 were identified to be unique norms based on healthy participants. The oldest norms on the RAVLT in this selection were published in 1989 and the newest in April 2023. Notably, only found two previous studies presenting norms based on a Scandinavian population.

First, the aforementioned study by Nielsen et al. (1989) also provided traditional norms for the RAVLT based on 101 participants between 20 and 54 years. However, the same limitations mentioned in relation to the TMT norms in section 1.5.2 also apply here.

Secondly, A. Vogel, Stokholm, and Jørgensen (2012) presented norms on the RAVLT based on a Danish sample of 100 participants. The norms were made using a traditional approach adjusted for age and education. Age was split in two levels (60-70, 71-78) and education in two levels (8-11 years, 12-17 years). The authors state that test scores in the Danish sample were similar to the published norms from the US (Strauss et al., 2006). Despite this, the authors argue for the need for more comprehensive norms on the RAVLT and point to the possibility of pooling data from several countries within Scandinavia in future studies like we did in Paper 2. Apparent disadvantages of these norms are the restrictive age span, and the small sample size in conjunction with the traditional norming methodology which requires

larger sample sizes for accurate norms (H. E. Oosterhuis et al., 2016; Van Breukelen & Vlaeyen, 2005).

The vast majority of recent normative studies report that performance on the RAVLT in healthy participants is significantly predicted by age, education, and sex (Alviarez-Schulze et al., 2022; Boenniger et al., 2021; Dassanayake, Hewawasam, Baminiwatta, Samarasekara, & Ariyasinghe, 2020; Kenyhercz, Fábíán, Andrejkovics, & Bugán, 2023; Lavoie et al., 2018; Ricci et al., 2022; Stricker et al., 2021; W. Van der Elst et al., 2017). Specifically, higher age, lower education, and female sex predicts better performance in adults and elderly.

1.5.6 Norms on RAVLT by Stricker et al. (2021)

In Paper 2 we assessed norms from Stricker et al. (2021) in a sample of Norwegians and Swedes. Previous studies have found partial support for the equivalence of US norms in Norway on another verbal learning test, the California Verbal Learning Test (CVLT) (Egeland et al., 2005), and US norms on the RAVLT have been suggested suitable in Denmark based on apparent similarities in mean and standard deviations for elderly (A. Vogel et al., 2012). Therefore, it was of interest to investigate how the US norms performed in our sample of Norwegians and Swedes. Specifically, the Stricker et al. (2021) norms were chosen because compared to previous norms used in Norway on the RAVLT (Schmidt, 1996), the Stricker et al. (2021) norms appear excellent in terms of sampling procedure, screening procedures, norming methodology, and reports an assuring sample size ($n = 4428$).

Stricker et al. (2021) reports that the norms were developed out of the Mayo Clinic Study of Aging (MCSA) which is a population-based aging study in Olmstead County, Minnesota, USA. They included participants from ages 30 to 89 years. Sampling was performed following a stratified approach for age and sex (i.e., sampling performed so that the final sample composition was balanced in terms of age and sex in ten-year intervals). Reportedly 60% of contacted participants in the population register for Olmstead County agreed to participate in the umbrella study (MCSA). Assessments in MCSA were performed from 2004 until March 2018.

According to Stricker et al. (2021), the total sample size for the RAVLT norms were 4428 participants and 98% of the sample is reported to be ‘white’. All participants performed the RAVLT for the first time. Assessments were performed by a psychometrician. For eligibility in the normative sample participants needed to be deemed ‘cognitively unimpaired’ by a physician and interviewing study coordinator. This was reportedly based on interviews and a

mental status exam and Clinical Dementia Rating (CDR). No cut-offs are reported on any measure, and participants were not excluded based on any other neuropsychological data. Thus, the normative sample of cognitively healthy adults in the Stricker et al. (2021) study sit somewhere in between what we might consider a pure population-based sample, and a thoroughly screened ‘undisputedly healthy’ sample.

The Stricker et al. (2021) norms were calculated using a regression-based approach. Raw scores were transformed to normally distributed scaled scores with a mean of 10 and standard deviation of 3 based on percentile ranks of raw scores. Scaled scores on basic and derived measures from the RAVLT were regressed upon pertinent demographical variables. Models were reported to fulfill the assumptions of linear regression analysis. There is no mention of any analysis for influential cases or outliers in Stricker et al. (2021). All basic RAVLT measures and Trials 1-5 total were adjusted for age, age², education, and sex. Age and education were entered as continuous variables. Years of education followed a structured coding scheme. Stricker et al. (2021) provide norms unadjusted for education, however, they recommended using fully adjusted scores adjusting for age, education, and sex whenever possible. In Paper 2 we applied the fully adjusted norms from Stricker et al. (2021) to calculate normed scores. Results from the Stricker et al. (2021) indicate significant effects of age, age², sex, and education on all primary subtests. Explained variance (R^2) is reported for combined models and individual predictors, however, they do not report partial R^2 (i.e., variance explained adjusted for the other predictors). Stricker et al. (2021) present test-retest reliability estimates for RAVLT raw scores based on a sample of 3555 participants tested with an average 1.4-year test-retest interval ($SD = 0.38$).

1.5.7 Color-Word Interference Test (CWIT) from the Delis-Kaplan Executive Function System (D-KEFS)

In Paper 3 we analyzed scores on the D-KEFS CWIT (D. C. Delis, Kaplan, & Kramer, 2001) which is a further development of the classical Stroop test first introduced in 1935 by John Ridley Stroop (Stroop, 1935). The CWIT is divided in four parts: Color-naming; color-reading; inhibition and inhibition/switching. The first three subtests are modeled after the original Stroop task. During the fourth and final subtest participants are required to alternate between color-reading and inhibition. This is unique to D-KEFS CWIT. Furthermore, a distinction between the D-KEFS CWIT and other versions of Stroop tests like Golden (Golden, Freshwater, & Golden, 1978) and Victoria (Regard, 1983) is that the primary

dependent measure in CWIT is time to completion measured in seconds for a fixed amount of items. Other versions of Stroop tasks often consider the number of correct responses in a fixed amount of time to be the dependent measure. An advantage of using time to completion as the dependent measure is that it eliminates ceiling effects.

The first two trials on the D-KEFS CWIT measures the basic abilities word reading and color naming. These tasks are thought to reflect focused attention and processing speed (Lezak et al., 2012). The third subtest inhibition also relies on the same basic abilities, but additionally requires the participant to inhibit the automated response of color-reading in favor of a more unfamiliar response. Inhibiting the automated response is demanding and the phenomenon by which participants take longer on this task is called the ‘Stroop interference effect’ and is by many considered the gold standard test for verbal response inhibition (MacLeod, 1992).

There is a large body of evidence correlating performance on Stroop tests with areas in the prefrontal lobes including the medial frontal gyrus, anterior cingulate cortex, and dorsolateral prefrontal cortex (Steinunn Adólfssdóttir et al., 2014; Duchek et al., 2013; Floden, Vallesi, & Stuss, 2011; Keifer & Tranel, 2013). Furthermore, clinical studies indicate that participants with lesions in the frontal lobes frequently perform worse on the task (Stuss et al., 2001). Also, impaired executive functions measured by Stroop tests in patients with amnesic MCI is a good marker for progression to dementia (Clark et al., 2012). Furthermore, performance on Stroop tasks is good for differentiating healthy participants from patients with dementia due to AD and correlates with degree of neuropathology (Bondi et al., 2002).

Delis et al. (2001) claim that the fourth CWIT subtest is the most difficult because it involves inhibition and cognitive flexibility, both of which are conceptualized as executive functions relying on frontal lobe functioning. Pilot data from Delis et al. (2001) indicate that the fourth trial was successful in separating clinical groups with frontal lobe dysfunction beyond what the conventional inhibition measure could. Subsequent studies have indicated the fourth subtest is not harder for all patients (Lippa & Davis, 2010), however, the measure may still have utilities in clinical groups characterized by deficits in cognitive flexibility or set shifting. That is, the task may be hard some clinical groups, but it not universally harder for all participants. Halleland, Haavik, and Lundervold (2012) showed that the fourth CWIT subtest managed to separate adults with Attention-Deficit/Hyperactivity Disorder (ADHD) from control participants, adjusted for differences in working memory and general cognitive ability. In contrast, the third CWIT task inhibition did not manage to separate groups.

1.5.8 Norms on D-KEFS CWIT and Stroop tests

To get an overview of the available normative studies on the D-KEFS CWIT I searched Pubmed for [(“Color-Word Interference Test”) and (D-KEFS) AND (normative)] which returned 4 results. None of the obtained results were norms, however there were some studies on the multivariate base rate of low or high performance on the entire D-KEFS battery adjusted for age and education (Karr, Garcia-Barrera, Holdnack, & Iverson, 2018).

To broaden the search and include other Stroop tests, I searched for [(“Stroop”) AND (normative)]. Out of the 138 returned results, 42 were identified to be unique norms based on healthy participants. Crucially, this only included written non-digital versions of Stroop tests (e.g., Victoria version, Golden Version). The results indicated that only two normative studies were based on Scandinavian samples.

First, Asmus Vogel, Stockholm, and Jørgensen (2013) developed norms on the Color Trails Test and the Modified Stroop Test (Klein, Ponds, Houx, & Jolles, 1997) based on Danish sample. In other words, they did not present norms on conventional Stroop paradigms.

Secondly, the previously mentioned study by Nielsen et al. (1989) also provided Stroop norms based on 101 Danish participants between 20 and 54 years of age using a traditional approach.

Virtually every recent normative study find that in samples of healthy participants, better performance on Stroop tests is predicted by lower age and higher education (Dassanayake, Hewawasam, Baminiwatta, & Ariyasinghe, 2021; Fábíán et al., 2023; Fällman et al., 2020; Ktaiche, Fares, & Abou-Abbas, 2022; Magnúsdóttir et al., 2021; Rodríguez-Lorenzana et al., 2021; Vicente et al., 2021). However, in these studies, sex differences are much less reliable. Some studies find that women outperform men, while others find no significant differences. Normative data on the CWIT specifically is scarce, however, longitudinal data suggest that scores on the CWIT decrease with age, but the slope over time does not vary between men and women, or levels of education (S. Adólfssdóttir, Wollschlaeger, Wehling, & Lundervold, 2017).

1.5.9 Age-adjusted norms on CWIT by Delis et al. (2001)

In lieu of local norms on the CWIT, the original age-adjusted norms from D-KEFS are frequently used by Norwegian clinicians and researchers (Ryder, 2021). One previous study has suggested that US norms from Delis et al. (2001) on the CWIT might be applicable in Norway based on apparent similarities between the normative estimates in a sample of adults

diagnosed with ADHD (Halleland et al., 2012), however, a more comprehensive analysis allowing for clearer recommendations based on a sample of healthy adults is warranted.

Data collection for the original age-adjusted norms from D-KEFS began in 1998 (Delis et al., 2001). The norms were based on 1750 participants between age 6-89 years. Delis et al. (2001) performed stratified sampling to ensure balance in the demographical composition of the sample and match the demographical composition in the 2000 US census. Participants were recruited via advertisements thus representing a self-selected convenience sample. The norms were developed using a regression-based approach adjusting for age with polynomials when pertinent. Raw score to age adjusted scaled score are reported in 10-year age intervals for adults (e.g., 20-29, 30-39 etc.). It is reported that each cell is based on approximately 75-175 participants. The D-KEFS manual states that norms adjusting for age, education, and sex were being developed in collaboration with Dr. Robert Heaton, however, we have not been successful in locating these norms on the CWIT, although more comprehensive D-KEFS norms have later on been developed for the TMT (Fine, Delis, & Holdnack, 2011).

2. Summary of Introduction

In summary, there is lacking normative data on popular tests in use in Norwegian and Scandinavian countries. Because of this, psychologists and researchers are using published norms from other countries – often from North America. However, previous studies indicate significant differences between normative estimates obtained in Scandinavia compared to the US. More generally, previous research indicates considerable heterogeneity between countries in terms of the average performance on cognitive tests, but also in the influence demographic variables have on scores. Norms are important because they form the basis for interpretation of individual scores. As such, inaccurate norms could have implications for the correct diagnosis and treatment for patients, as well as stratification and selection for research. On the measures TMT, RAVLT, and CWIT there is a lack of normative data available from Scandinavian countries and there is a need for new norms based on local samples.

3. Objectives

The primary objective of this thesis was to develop regression-based norms on three popular neuropsychological tests used in Norway and Scandinavia, namely the Trail Making Test, Rey Auditory Verbal Learning Test (RAVLT) and Color-Word Interference Test (CWIT).

Secondly, in Papers 1-3 we aimed to compare the new Scandinavian norms with published norms frequently used by Norwegian psychologists. We assessed whether these previously

published norms adequately adjusted for demographical variables in our Scandinavian samples and generally align with the expected cognitive performance in these samples (i.e., if the resulting *T*-scores significantly differ). In this regard, a general research question is whether there can be observed consistent differences between norms based on Scandinavian samples and published norms from North America, and whether these North American norms are suitable in Scandinavian samples.

4. Methods and materials

4.1 Background on the cohorts

This thesis is based on cross-sectional data from healthy participants sourced from four cohorts: The dementia disease initiation study (DDI) (Paper 1, 2, 3); Gothenburg MCI (Paper 1 & 2); Center for Lifespan Changes in Brain and Cognition (LCBC) (Paper 3) and Oslo MCI (Paper 2 & 3). Longitudinal data from one follow-up was supplied from DDI, Gothenburg MCI and LCBC for test-retest analysis in Papers 2 & 3. To familiarize the reader on the research context the following sections provide a summary on the aims and methods of the cohorts.

The DDI study is a longitudinal multicenter study on preclinical and prodromal phases of common dementia diseases including but not limited to AD (Fladby et al., 2017).

Assessments in DDI started in January 2013 and is still ongoing. The DDI study includes participants between age 40-80 years. The study recruits control participants which consist of normal healthy adults to contrast symptom group participants in preclinical and prodromal phases of AD, or participants with subjective cognitive decline (SCD). Assessments in DDI were conducted at hospitals throughout Norway: Akershus University Hospital; University Hospital of North Norway in Tromsø; Betanien Hospital in Bergen; St. Olav's Hospital in Trondheim; Haugesund Hospital and Stavanger University Hospital. The DDI study incorporates an extensive neuropsychological test battery and a vast array of biomedical measures. Oslo MCI (Hessen et al., 2014) is the predecessor of the DDI study. As such, the core study protocols, aims, and procedures in Oslo MCI and DDI are similar. Assessments in Oslo MCI were performed between 2004 and 2011.

Gothenburg MCI is a longitudinal single-center study on common dementia diseases and vascular pathology similar to DDI in scope and breadth (Wallin et al., 2016). Assessments in

Gothenburg MCI were conducted at Sahlgrenska University hospital in Gothenburg Sweden between January 2001 and March 2014. Like DDI, Gothenburg MCI employs a multi-modal approach involving several professional groups in addition to a broad neuropsychological test battery.

LCBC is a multi-disciplinary research center at the Department of Psychology, University of Oslo, aimed at investigating longitudinal trajectories in cognition and brain health in healthy participants and patients with neurodegenerative diseases like AD (Fjell et al., 2018). LCBC consists of several sub-studies and for this thesis healthy participants between 20 and 85 years were sourced from the following: Neurocognitive development (Tamnes et al., 2013); Neurocognitive plasticity (de Lange, Bråthen, Rohani, Fjell, & Walhovd, 2018) and Biological predictors of memory (Storsve et al., 2014). LCBC also incorporates a multi-faceted approach in their studies including magnetic resonance imaging (MRI), blood, electroencephalogram (EEG), neuropsychology, transcranial direct current stimulation (tDCS) and virtual reality paradigms. As a note, LCBC is considered a center of excellence and has been named a world leading research group within their research fields by the Norwegian Ministry of Education (Milde, 2015).

4.2 Study samples, joint inclusion and exclusion criteria, and recruitment methods

Participants recruited for Papers 1-3 were interviewed to assess their cognitive and somatic health status. All participants enrolled in the main normative analyses in Papers 1-3 were considered healthy participants. For Papers 1 & 2 inclusion criterion were age 40-80 years. For Paper 3 inclusion criterion were age 20-85 years. All participants were either native Norwegian (Paper 1-3) or Swedish (Paper 1 & 2). Almost all participants in Papers 1-3 were of European ethnicity.

Healthy participants were excluded if they reported a history of severe psychiatric or somatic disease that might influence cognitive functions, brain trauma, dementia, severe learning disabilities, or developmental disorders. Furthermore, the self-reported experience of subjective cognitive decline (SCD) was an exclusion criterion for participants in the main normative analyses in Paper 1-3. Criteria employed for defining SCD is detailed in section 6.3. Participants' cognitive normalcy was screened with the MMSE (Folstein, Folstein, & McHugh, 1975). In Paper 1 & 2 we set a cut-off criterion of ≥ 26 on the MMSE for participation as healthy participants. In Paper 3 we used the MMSE for screening cognitive normalcy, but we did not employ a strict cut-off for selecting participants. Demographical variables and scores on MMSE for samples in Papers 1-3 are detailed in Table 1.

Psychological symptom scales including the GDS-15 and GDS-30 (Yesavage et al., 1982; Yesavage & Sheikh, 1986), Symptom Check List 90 (Derogatis & Unger, 2010) and Beck Depression Inventory (Beck, Steer, & Brown, 1987) were used during clinical interviews to assist in the clinical assessment of psychiatric symptoms. However, we did not define cut-off criteria on these scales for participation as a healthy participant due to incomplete records and missing values.

Healthy participants in the normative analyses in Papers 1-3 from DDI and Oslo MCI were predominately spouses of symptom group participants, volunteers responding to advertisements in news outlets, or patients recruited from an orthopedic ward at Akershus University Hospital. Healthy participants from LCBC in Paper 3 were predominately recruited through local workplaces and Universities in the greater Oslo region, or as volunteers responding to advertisements. Participants from Gothenburg MCI in Paper 1 & 2 were mostly recruited through senior citizen organizations, and some were recruited via relatives already participating in the study.

4.2.1 Independent Comparison Group for Paper 2

145 cognitively healthy participants that self-reported an experience of SCD from DDI and Gothenburg MCI served as an independent comparison group to assess the Scandinavian norms and previously published norms from the US. SCD was classified during a standardized protocol based on standardized criteria (Jessen et al., 2014). A clinical interview was conducted with all participants to assess what specific domains participants experienced decline in, the nature of progression since onset, familiar history, and affective symptoms. All participants reporting SCD performed within expected ranges for cognitive normalcy on screening tests, and otherwise fulfilled the criteria for participation as healthy participants previously described. Following criteria from Albert et al. (2011) participants reporting SCD were excluded if they obtained scores more than 1.5 *SD* below the normative mean on at least one of the following neuropsychological tests (test reference and normative reference in parenthesis): Silhouettes from Visual Object and Space Perception Battery (Eliassen et al., 2020; Warrington, 1991); Controlled Oral Word Association Test (Heaton, 2004; Lorentzen et al., 2023) and Trail Making Test part B (Espenes et al., 2020; Reitan & Wolfson, 1985). Participants from DDI and Gothenburg MCI were assessed using the same tests, except participants from DDI were tested with the subtest Delayed Word List Recall from the CERAD battery (Fillenbaum et al., 2008; Kirsebom et al., 2019), while participants from Gothenburg MCI instead were tested with the thirty-minute delayed recall trial from RAVLT

(Rey, 1958; Stricker et al., 2021). Cognitively healthy participants with SCD in the independent comparison group were included via referral from general practitioners to memory clinics, or self-referral responding to advertisements in local newspapers and media aimed at including participants with memory complaints.

5.2.2 Test-retest samples in Paper 2 & 3

98 participants from DDI and Gothenburg MCI had available data from one follow-up on the RAVLT (Table 1). All participants in the follow-up sample fulfilled eligibility criteria for study participation as healthy participants on baseline and follow-up. None of the participants progressed to MCI, dementia, or self-reported SCD at follow-up. The average test-retest interval was 2.6 years ($SD = 0.5$). For Paper 3, 335 participants from LCBC had available data from one follow-up on the CWIT (Table 1). All participants in the sample fulfilled inclusion criteria and none of the exclusion criteria at baseline. Eligibility criteria were not assessed at follow-up. To keep the test-retest interval more homogenous we excluded participants tested later than 5 years after follow-up ($n = 22$). As a result, the average test-retest interval was 3.4 years ($SD = 0.9$).

Table 1

Demographical variables and descriptive statistics for samples in Papers 1-3

		Age $M (SD)$ [range]	Edu $M (SD)$ [range]	Female $n (%)$
Paper 1	Healthy participants ($n = 292$)	63 (8.4) [41 – 84]	13.2 (3.4) [6 – 24]	174 (59.6%)
Paper 2	Healthy participants ($n = 244$)	64.3 (6.8) [49 – 79]	12.7 (3.3) [6 – 24]	138 (56.6%)
	Comparison group ($n = 145$)	62.3 (6.7) [49 – 77]	14 (3.2) [6 – 21]	91 (62.8%)
	Test-retest sample ¹ ($n = 98$)	63.9 (6.7) [49 – 77]	12.5 (3.2) [6 – 24]	65 (66.3%)
Paper 3	Healthy participants ($n = 1011$)	46.2 (19.1) [20 – 85]	15.5 (2.9) [7 – 23]	675 (66.8%)
	Test-retest sample ¹ ($n = 335$)	52.7 (18.4) [20 – 84]	15.6 (2.9) [8 – 23]	207 (61.8%)

Note. ¹Age and education at baseline; Edu = years of education

4.3 Materials

Paper 1: Trail Making Test (Reitan & Wolfson, 1985)

Administration procedures and test instructions on the TMT are described in the methods-section of Paper 1 (Espenes et al., 2020). The dependent measure on TMT-A and TMT-B is time to completion measured in seconds. The outcomes on TMT can be divided into basic and derived measures. Basic measures are time to completion on TMT-A and TMT-B, and derived measures are those which are calculated from the scores on TMT-A and TMT-B.

Derived measures are calculated in an attempt to isolate the additional task demands associated with TMT-B, namely working memory and cognitive flexibility. Similarly, TMT- β is a derived measure that is computed by analyzing scores on TMT-B regressed on predictors age, education, and scores on TMT-A (Table 2).

Paper 2: Rey Auditory Verbal Learning Test (RAVLT) (Rey, 1958)

Administration procedures and test instructions on the RAVLT are provided in the appendix of Paper 2 (Espenes et al., 2022). Materials on the RAVLT in Paper 2 were translated and provided to the DDI study by Professor Erik Hassen at the University of Oslo, however, as far as we are aware this material is not translated according to standardized procedures (A. Evers et al., 2013). Likewise, the Swedish translations were provided to us by psychologists in Gothenburg MCI at the Sahlgrenska University Hospital and it is unclear whether best practice guidelines for adaptation were followed (A. Evers et al., 2013).

On the RAVLT, the dependent measure is the number of correctly recalled words by the participant for each trial. The outcomes can be divided into basic and derived measures which attempt to isolate different processes in episodic memory. All basic and derived measures analyzed in Paper 2 are presented in Table 2.

Paper 3: D-KEFS CWIT

The CWIT was administered according to standard instructions described in the Norwegian D-KEFS manual supplement (D. Delis, 2005) and Paper 3. All test stimuli were in Norwegian, provided by Pearson Assessment. On the CWIT the outcome is time to completion of a fixed number of items (Table 2).

Table 2*Primary and derived measures on TMT, RAVLT and CWIT*

<u>Trail Making Test (TMT)</u>	
TMT-A	Time to completion connecting numbers
TMT-B	Time to completion alternating numbers and letters
TMT B-A	Score on TMT-B minus score on TMT-A
TMT B/A	Score on TMT-B divided by score on TMT-A
TMT-β	TMT-B ~ Age + Education + TMT-A ¹
<u>Rey Auditory Verbal Learning Test (RAVLT)</u>	
	# Correctly recalled words from list A directly after presentation
Trial 2	Second learning trial
Trial 3	Third learning trial
Trial 4	Fourth learning trial
Trial 5	Fifth learning trial
	# Correctly recalled words from list B directly after presentation
List B	
Trial 6	Recall of list A without renewed presentation
Trial 7	Recall of list A after thirty minutes without renewed presentation
Trials 1-5 total learning (TL)	\sum (Trial 1, Trial 2, Trial 3, Trial 4, Trial 5)
Learning over trials (LOT)	\sum (Trials 1–5 total - (Trial 1*5))
Learning rate (LR)	\sum (Trial 5—Trial 1)
Proactive inhibition	\sum (Trial 1—list B)
Retroactive inhibition	\sum (Trial 5—Trial 6)
Long-term percentage retention (LTPR)	\sum (100 * (Trial 7/Trial 5))
<u>Color-Word Interference Test (CWIT)</u>	
CWIT-1	Time to completion color-naming
CWIT-2	Time to completion color-reading
CWIT-3	Time to completion inhibition
CWIT-4	Time to completion inhibition/switching

Note. \sum = sum; ¹ ~ = regressed on. # = “number of”. Primary measures are listed in the order they are administered.

4.4 General procedures

All cohorts featured in this thesis followed a similar procedure starting with interviews to record health information from participants. During this interview, information on the educational background of the participants was recorded. Education was encoded as total years of education rounded down to the nearest whole number. Every year of formal education attained was counted, excluding degrees or schooling of the same level, or further

education obtained for instance through a professional position. For example, a participant reporting 12 years of basic schooling, a three-year bachelor's degree, and two separate master's degrees each normed for two years, was only recorded as having 17 years of education in total.

Participants in Papers 1-3 were further examined during physical assessments by a medical doctor or neurologist. Depending on their participation status they were further referred to MRI, blood tests, cerebrospinal fluid (CSF) and other examinations. Results from these were not analyzed in any analyses in the current thesis. The neuropsychological assessments were conducted by either psychologists, psychologists-in-training, medical doctors, or study nurses. All administrators received training by qualified professionals such as psychologists specializing in neuropsychology and were supplied instructions on standardized administration and scoring of tests. Pertinent test translations and procedures are detailed in the method sections of Papers 1-3. The neuropsychological measures analyzed in this thesis was part of a broader test battery that differed between cohorts. For information on test batteries and procedures for each cohort, please refer to previous publications (Fjell et al., 2018; Fladby et al., 2017; Wallin et al., 2016).

4.5 Statistical Analyses

4.5.1 Comparisons of mean scores

Throughout Papers 1-3, we assessed mean differences in raw scores and *T*-scores with independent samples t-tests (or Welch's t-test in case of unequal variances) or Mann-Whitney U test for non-normally distributed variables (James et al., 2013). Additionally, for the purpose of this thesis we conducted two-tailed t-tests to obtain Cohen's *d* effect size estimates for mean differences between men and women on Trial 7 of the RAVLT not reported in Paper 2 (section 5.2 Summary of results Paper 2).

4.5.2 Normalization procedures Paper 1 and 3

For Paper 1 and 3, the dependent variable is time to completion and distributions were positively skewed. We therefore retrieved the cumulative distribution and transformed raw scores to scaled scores ($M = 10$, $SD = 3$). For instance, the 50th percentile for a raw score corresponded to scaled score 10. Raw score to scaled score conversions were reversed so that longer time to completion on the TMT and CWIT corresponded to a lower scaled score. For instance, the 99th percentile corresponded to scaled score 3. In Paper 1 this was done manually by retrieving the raw score to percentile distributions via the Frequency function in Statistical

Package for Social Sciences (SPSS) version 25. In Paper 3 we retrieved the reverse cumulative frequency of raw scores on the CWIT using the package Classical Test Theory Functions (CTT) in R studio version 4.2.1 (Willse, 2022). For Paper 2 there was no need to transform dependent measures from the RAVLT as raw scores, and regression residuals were adequately normally distributed.

4.5.3 Multiple regression analysis Papers 1-3

Predictor selection. For selecting predictors in the normative analyses, we conducted preliminary analyses correlating age and education with scores on pertinent measures using Pearson's R or Spearman's Rho depending on the properties of the pertinent cognitive outcome. We then included all terms assumed to relate to test performance based on previous studies or found to relate to performance in the preliminary analyses. These were included as independent variables in multiple regression analysis with the pertinent cognitive outcome as dependent variable. Typically, this included age, education, sex, in addition to polynomial terms (e.g., age² and age³) and interactions (e.g., age*education).

We then performed series of regression analyses and hierarchically dropped terms in a stepwise manner following a-priori set statistical criteria ($\alpha = .01$). Compared to other methods for predictor selection like simultaneous regression, backwards regression analyses may increase the family wise error rate (inflate type 1 error) because there is no way to adjust for multiple testing (Hastie, Tibshirani, Friedman, & Friedman, 2009; H. Oosterhuis, 2017). However, we took great care in assessing predictors not only based on p -values, but also the explained variance by predictors (partial R^2), ANOVAs of nested models, and Bayesian Information Criterion (BIC), thereby reducing the risk for chance capitalization and by extension norming scores according to predictors that were not relevant for task performance. Furthermore, predictor selection was conducted independently by at least two authors in Papers 1-3. Any discordance in model structure obtained by authors were resolved by discussion considering the obtained p -values, BIC, adjusted R^2 , omnibus ANOVAs and previous studies.

Assessment of influential cases and outliers. An outlier is defined as an observation with a predicted score that varies greatly from the observed scores (i.e., an observation with a large residual value). Studentized residuals of ± 3 indicates an outlier, however, depending on the sample size residuals slightly larger than this might be expected. Outliers are typically easily spotted on plots of fitted values versus residuals. Outliers do not necessarily heavily influence the regression estimates but can cause issues for the residual standard error, thereby

influencing the p -values, confidence intervals, and total explained variance of the model R^2 (James et al., 2013).

In Papers 1-3, an influential case was defined either by leverage values or Cook's D. Leverage is a measure of whether the observations have extreme predictor values. Cook's D is a measure of how much the regression equation changes with the observation deleted (James et al., 2013). Of course, influential cases are cause for concern in normative studies because few observations may skew the joint prediction when the norms are in fact intended to model the average association between demographic variables and cognitive test scores. In Papers 1-3 we assessed plots of standardized residuals and leverage / Cook's D. We did not follow a priori statistical criteria to define concerning Cook's D values, but instead relied on visual analysis of plots to detect observations that deviated greatly.

Assumptions. After obtaining an adequate model structure we assessed whether the final models fulfilled assumptions for linear regression analysis. The regression analyses performed in the samples are conducted to estimate the true unknown relationship between the regression predictors and the outcome in the population. However, the validity of the sample analysis to accurately describe the unknown true relationship in the population relies on certain assumptions about the predictors, the outcome, and the model specification. These assumptions are called 'Gauss-Markov assumptions' and if these are met then the regression analysis is said to be the best linear unbiased estimator of the true relationship between the predictors and the outcome in the population (Berry, 1993, pp. 18-19, as cited in Oosterhuis, 2017). Thus, the assumptions are crucial for valid inference.

Homoscedasticity. The residuals describe the difference between the predicted score for an individual and the actual obtained score. Homoscedasticity refers to the constant variance of the residuals across all values of the covariates. In regression-based norming, homoscedasticity is an important assumption because the SD of the residual is used for calculating normative scores and the variance is assumed to be constant regardless of the predicted score. Thus, if the residuals are heteroscedastic there is an increased risk for faulty estimation of the norms. If residuals are found to be heteroscedastic for some area of the fitted values, the SD of the residual can be calculated separately for each quartile of the fitted values (Wim Van der Elst et al., 2006). This is important because if the true residuals for a given value of the predictor is larger than the residual estimate then the norms will be underestimated. Reversely, if the true residuals are smaller than the estimate the norms will overestimate (H. Oosterhuis, 2017). For example, if the predicted score is 8.13 and the

obtained score is 7, and the *SD* of the residual is estimated to be 2.775, the *Z*-score is calculated to 0.41 $((8.13 - 7)/2.775)$ which equates to $T = 46$. But if the true variance for this predictor value were lower, say 2.2, then the *Z*-score is 0.51 ($T = 49$). In addition, homoscedasticity is an important assumption because the residual standard error is used for hypothesis testing of coefficients and associated confidence intervals. Thus, heteroscedasticity will bias the power of the analyses (James et al., 2013).

Linearity and additivity. Additivity means that the relationship between predictor 1 and the outcome (Y) does not vary according to levels of predictor 2. Linearity means that the association between predictor 1 and Y is constant for all levels of predictor 1 (James et al., 2013). If the true relationship between the predictors and the outcome is not linear, then the conclusions drawn from the model is ‘suspect’ and prediction using the model is not accurate (James et al., 2013). This may be especially important in normative studies where the model predictions are used to make important inferences. Models with included polynomial- or interaction terms are still considered linear because they use a linear function to fit the data and are said to be intrinsically linear regression models (Nisbet, Elder, & Miner, 2009; H. Oosterhuis, 2017). For the assumptions of linearity and additivity we always assessed whether polynomial or interaction terms improved model fit. We added polynomials or interactions if *p*-values, ANOVAs, BIC and R^2 indicated that it was necessary in the model selection phase. Furthermore, we assessed plots of standardized residuals vs. fitted values for all regression equations (James et al., 2013). Plots were fitted with a line indicating a smooth fit and no discernable patterns or trends indicated homoscedasticity. Both homoscedasticity and linearity/additivity were assessed with residual plots using base R functions.

Collinearity. Collinearity refers to when predictors are highly correlated. This is problematic because collinearity increases the standard error (*SE*) of the coefficients thereby skewing *p*-values and confidence intervals. We checked this by correlating the predictors and mean-centering predictors before analysis. Collinearity was assessed with variance inflation factor that did not exceed 5 for any analysis (James et al., 2013).

Normal distribution of residuals. For most applications, outcomes used in linear regression do not need to be normally distributed and it is the normal distributions of residuals that is essential (James et al., 2013; Lumley, Diehr, Emerson, & Chen, 2002). In regression-based norming the residuals are used to calculate normed scores, and the residuals are interpreted via a *Z*-score distribution (Wim Van der Elst et al., 2006). Thus, it is important that the residuals are normally distributed for valid norms. As is customary, we visually

assessed the normality of residuals and dependent measures with histograms and QQ-plots (James et al., 2013).

Independence of residuals. The regression models assumes that each observation is independent, i.e., uncorrelated, with other observations. Typically, violation of this assumption is observed in longitudinal data where the same individual is measures multiple times, time-series data, or data that is clustered in some other way (James et al., 2013; W. Van der Elst et al., 2017). Correlated errors can artificially lower the *SEs* and result in false positive results because the correlation between observations is not accounted for. We assessed the potential for correlated errors in the design phase of the analyses and did not perform any multi-level analyses to challenge the assumed independence of observations.

4.5.4 Multivariate regression analysis Paper 2

In Paper 2 we followed procedures described in W. Van der Elst et al. (2017) for multiple multivariate regression-based norming. First, we assessed associations between RAVLT subtests and demographic variables with Pearson correlations to determine how related subtests were and identify likely predictors for the normative regression models. For the model selection phase, we started with a full model including all predictors assumed to relate to performance on the RAVLT based on previous studies and the correlation analysis. The full model can be summarized as: [RAVLT scores ~ Age + Age² + Edu + Edu² + Age*Edu + Age*Sex + Trial + Trial*Age + Trial*Edu + Trial*Sex]. Models were fitted using the `gls()` function in the package “nlme” in R (Pinheiro et al., 2017). Age and years of education were mean centered to avoid issues with multicollinearity and improve interpretation of the intercept. Sex was coded as female = 1 and male = 0. We tested interaction terms for whether the effect of demographic variables significantly differ across subtests on the RAVLT.

The model selection procedure followed a backwards selection procedure. In a stepwise manner, we dropped non-significant terms one at a time until the model structure could not be simplified further without deterioration (James et al., 2013). The backwards selection was guided by -2 log-likelihood ratio of models, Bayesian information criterion (BIC) and Akaike information criterion (AIC) (W. Van der Elst et al., 2017). -2 Log-likelihood ratios were compared with the `anova()` function in R. If the *p*-value exceeded the a priori set alpha level criterion of $\alpha = .01$ this indicated that the simplified model did not explain significantly less variance in scores. As a result, this model would be used as reference for further simplification. The backwards selection continued until *p*-values dropped below .01,

indicating that the model could not be simplified any further. The final model can be summarized as: [RAVLT scores ~ Age + Edu + Sex + Trial + Trial*Edu + Trial*Sex].

By default, the multivariate model is fitted with an unstructured residual structure, but simplification of this structure could potentially increase the power of the analysis by estimating the variance-covariance structure (W. Van der Elst et al., 2017). This reduces the estimated parameters in the model and as a result less degrees of freedom are used. As described in W. Van der Elst et al. (2017), we therefore fitted models with homogenous and heterogenous auto-regressive and compound symmetry structure. However, this did not retain model fit to a satisfactory degree. Finally, as recommended in W. Van der Elst et al. (2017) we re-fitted the final model using restricted maximum likelihood (REML) instead of the default maximum likelihood (ML) as this may provide better estimates of the variance parameters such as the *SD* residual (Verbeke & Molenberghs, 2009). The assumptions of heteroscedasticity and normality were assessed using plots of standardized residuals and fitted values. Plots indicated no homoscedasticity or trends indicative of non-linearity or additivity. No apparent outliers were visible on plots (standardized residual minimum = -3.3, maximum = 3.3). The final standardized residuals were normally distributed, which was supported by QQ-plots also indicating that residuals were approximately normally distributed.

4.5.5 Calculating normative performance

The method for calculating the norms was similar in all three Papers and were based on methods described in previous studies (Kirsebom et al., 2019; Parmenter et al., 2010; Wim Van der Elst et al., 2006). While the exact method for calculating the predicted scores differ in Papers 1-3, the regression-norms are in essence calculated by the following simple formulae: $[(\text{Obtained score} - \text{Predicted score}) / (\text{SD of the residuals})] = Z\text{-score}$. The *Z*-score can be further converted to a *T*-score ($M = 50$, $SD = 10$) by: $[(Z\text{-score} * 10) + 50]$.

In Papers 1-3 we also provide html calculators that compute the regression equations based on inputs of raw scores and pertinent demographical variables to provide quick and easy estimations of norms on all measures. This significantly reduces the computational demands for users of the norms and might reduce occurrence of computation error. Links to the calculators are provided in the method section of each paper.

4.5.6 Assessment of published norms in Scandinavian samples

Assessing the published norms in the Scandinavian samples was done following these strategies: Firstly, we computed *T*-scores adjusted for pertinent demographics using published

norms and our newly developed norms on outcomes of the cognitive tests. *T*-scores were calculated following procedures previously described, or as described in the published norms. As a result, for all participants we had paired sets of *T*-scores that were calculated either with published norms or our own norms on all outcomes of interest.

The comparison and assessment of norms was then conducted as: 1) Multiple regression analyses with the *T*-scores as dependent variables and age, education, and sex as predictors. The rationale behind these analyses were that *T*-scores are expected to be adjusted for demographical variables and if norms are sufficiently adjusted for demographical variables, then the predictors should not explain variance in *T*-scores. Similar approaches are reported in previous publications (Hestad et al., 2016). For instance, while higher age is associated with lower scores on the RAVLT, norms should adjust for this, and elderly participants should not on average receive lower *T*-scores than young participants. Thus, age should not explain variance in age-adjusted *T*-scores. In Paper 2 we first assessed omnibus ANOVAs for overall significance with an alpha level criterion of $\alpha = .01$ because a non-significant result indicates that none of the included predictors significantly predict the outcome (James et al., 2013). However, if omnibus ANOVAs were significant, we then analyzed models for adjusted R^2 and individual predictors for partial r^2 , p -value, and the direction of the unstandardized beta coefficients following a criterion of $\alpha = .05$. In Paper 1 and Paper 3 we analyzed p -values from individual coefficients following a conventional alpha level criterion of $\alpha = .05$.

2) We assessed distributions of *T*-scores for overall *M* and *SD* and compared norms using paired samples t-tests for mean differences. In Paper 1 we compared mean differences in *T*-scores split by high and low education to highlight differences between norms. In Paper 2 and 3 we instead plotted *T*-scores and predictors in a scatterplot with a linear fit to indicate trends of adjustment across the entire range of predictors. To illustrate this, similar plots were created for this thesis that show the *T*-scores and demographic predictors on the TMT. Lastly, for Paper 3 we compared the percentage of participants in the normative sample that were identified as obtaining a *T*-score ≤ 35 using our norms and published norms. We compared observed rates to the expected base rate (6.7%) using two-tailed one proportion Z-tests and calculated the 99% *CI*s around the estimate. Lastly, we compared whether the observed rates significantly differed between norms with paired samples proportion tests.

4.5.7 Percentiles on skewed measures

In case the assumptions of linear regression were not met for any cognitive outcome we instead calculated percentiles based on the observed cumulative distribution in the sample.

This was either conducted as obtaining percentiles associated with observed raw scores, or estimating the raw score associated with specific percentiles. Typically, the percentile approach is used if cognitive outcomes or residuals from regression analysis were non-normally distributed. For instance, error measures on cognitive tests are often zero-inflated and over-dispersed which makes error measures unsuited for general linear regression-based norming. To determine the need for stratification according to demographic variables we assessed linear associations between cognitive outcomes and demographic variables either through Pearson's correlations or independent samples T-tests (or pertinent non-parametric alternatives). If the outcomes were significantly associated with demographic variables, we split the sample according to relevant variables and calculated the percentiles in discrete distributions (e.g., men and women separately).

4.5.8 Test-retest reliability

We provide test-retest reliability indices from participants with available follow-up measures on the RAVLT and the CWIT (Paper 2 and 3). For Paper 2 and 3, we calculated intraclass correlations (ICC) on the RAVLT and CWIT between baseline *T*-scores and follow-up. *T*-scores at follow-up was calculated with updated age and education if applicable. ICCs and 95% confidence intervals (*CI*s) in Paper 2 & 3 were calculated based on single rating, absolute-agreement two-way mixed-effects models (McGraw & Wong, 1996). Compared to other common methods like Pearson's correlation, the ICC is preferred for estimating test-retest reliability because it reflects both degree of correlation and agreement (Koo & Li, 2016). As a result, in some instances Pearson's correlation might indicate a better test-retest reliability than is warranted. For example, if all participants consistently scored +3 *T*-scores higher from baseline to follow-up, Pearson's *R* would indicate a perfect correlation of 1.0. In contrast, the ICC would correctly describe the lower agreement between baseline and follow-up (ICC = .787) (Koo & Li, 2016). ICC analyses were done in R studio using the function 'ICC ()' from the package "psych" (Revelle & Revelle, 2015).

4.6 Ethics

Regional medical research ethics committees in Norway and Sweden approved the underlying studies this thesis draws upon. All studies were conducted in accordance with the Helsinki declaration of 1964 (revised 2013). The Norwegian Health and Research Act were followed regarding data storage and privacy concerns. Patients with dementia or MCI may be considered an especially vulnerable group where special consideration when conducting research is warranted (Slaughter, Cole, Jennings, & Reimer, 2007). However, dementia or

MCI were cause for exclusion in Papers 1-3, and in the studies the present thesis draws on dementia was an exclusion criterion for participation. All participants were informed on their right to withdraw from participation, potential risks and benefits associated with participation, study procedure and privacy, and signed written informed consent sheets.

5. Summary of results

5.1 Summary results Paper 1

Effect of demographics on basic and derived TMT measures.

Higher age was on average associated with significantly lower scores on TMT-A (partial $r^2 = .17$, $p = <.001$). On TMT-B, higher age and lower education were significantly associated with lower scores (partial $r^2 = .159$, $p = <.001$, partial $r^2 = .036$, $p = <.001$, respectively).

Higher age and lower education were significantly associated with lower scores on the derived measure TMT B/A (partial $r^2 = .069$, $p = <.001$, partial $r^2 = .042$, $p = <.001$, respectively). On TMT B/A, lower education was significantly associated with lower scores (partial $r^2 = .025$, $p = <.01$).

On our proposed measure TMT- β , higher age (partial $r^2 = .055$, $p = <.001$), lower education (partial $r^2 = .042$, $p = <.001$), and lower TMT-A scaled scores (partial $r^2 = .209$, $p = <.001$) were associated with decreased scores on TMT- β . Correlation analysis indicated that TMT- β was not significantly associated with TMT-A scores (i.e., TMT- β successfully adjusted for TMT-A performance) ($r = -.003$). TMT- β and TMT B-A were highly correlated ($r = .969$, $p = <.001$).

Comparison with published norms from Heaton (2004) and Tombaugh (2004).

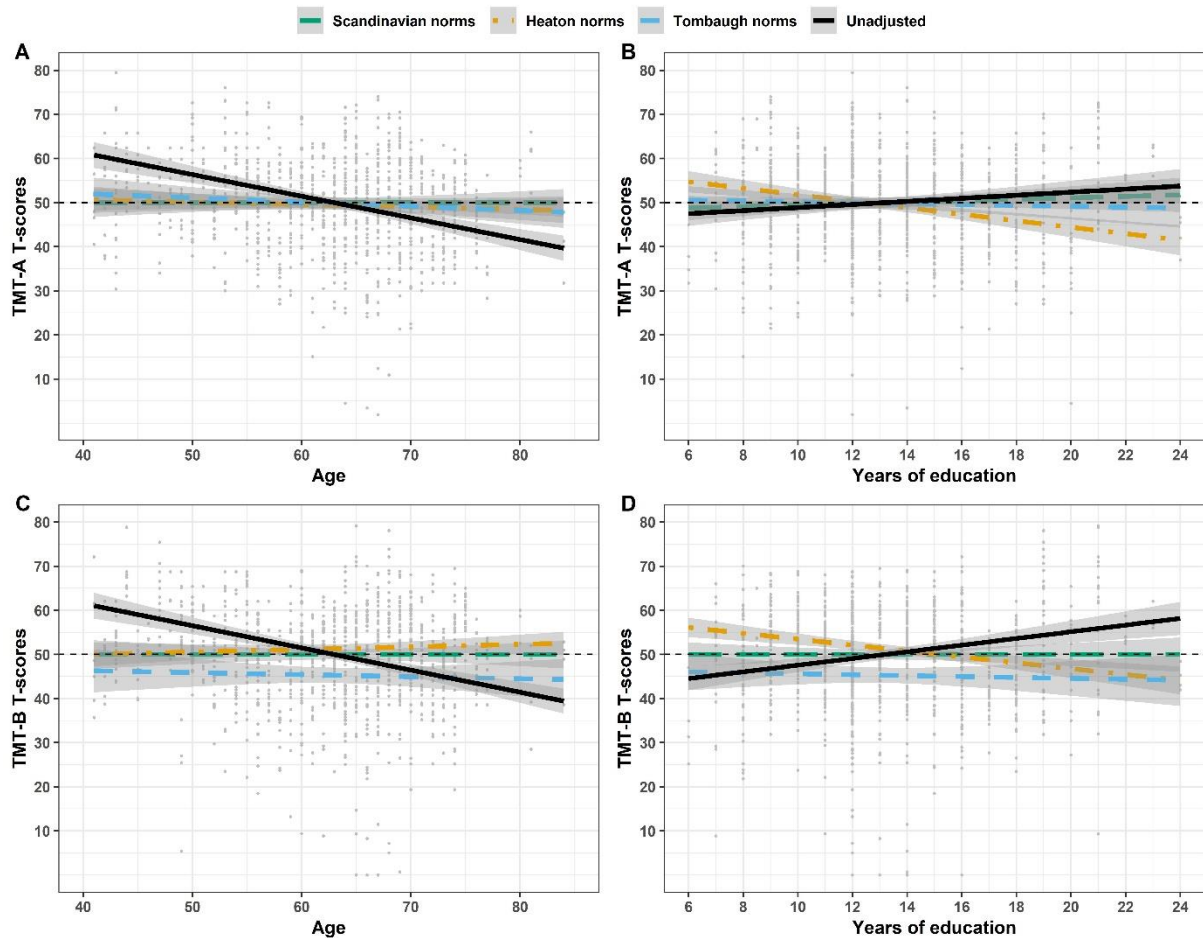
Norms from Heaton (2004) did not adequately adjust for the effects of education in the Scandinavian normative sample ($n = 292$) on TMT-A or TMT-B. On average, there was a linear trend where higher educational attainment predicted lower T -scores on TMT-A ($b = -0.771$, partial $r^2 = .079$, $p = <.001$) and TMT-B ($b = -0.661$, partial $r^2 = .068$, $p = <.001$). Further analysis indicated that this was because participants with low educational attainment were estimated significantly higher T -scores compared to our local norms ($> T50$) and participants with high educational attainment were estimated significantly lower T -scores ($< T50$) (Figure 1). Norms from Heaton (2004) were too lenient for participants with 12 or less

years of education on TMT-A ($M = 51.85$, $SD = 9.08$) and TMT-B ($M = 53.18$, $SD = 7.9$). For these participants, the differences with our local norms were significant on TMT-A ($t(142) = 6.74$, $p < .001$) and TMT-B ($t(142) = 9.78$, $p < .001$). Reversely, norms from Heaton (2004) were too strict for participants with 13 or more years of education on TMT-A ($M = 46.95$, $SD = 8.77$) and TMT-B ($M = 49.35$, $SD = 8.96$). Again, for these participants, the mean differences were significant compared to our local norms on TMT-A ($t(123) = -12.32$, $p < .001$) and TMT-B ($t(123) = -2.41$, $p = .017$).

Norms from Tombaugh (2004) successfully adjusted for the effects of age and education in the sample (i.e., there were no significant linear trends). Scores on TMT-A were close to the expected normative mean ($M = 49.86$, $SD = 11.51$) and compared to our local norms the mean difference in T -scores was not significant ($t(291) = 0.41$, $p = .678$). However, participants from Scandinavia were on average estimated to have approximately 0.5 SD lower T -scores on TMT-B ($M = 45.3$, $SD = 15.1$) which was significantly lower compared to our local norms ($t(291) = 9.26$, $p < .001$). Furthermore, the variance in scores was high and visual inspection of the distributions indicated considerable negative skewness and kurtosis on both TMT-A and TMT-B.

Figure 1

Linear plots of TMT T-scores computed with norms from Heaton (2004), Tombaugh (2004), unadjusted scores, and our local norms from Scandinavia



Note. Reference line indicates the expected normative performance of approximately T 50; Figure not included in Paper 1 and was constructed for illustration in this thesis.

5.2 Summary results Paper 2

Effect of demographics on RAVLT measures.

Higher age and lower education were associated with lower scores on all primary RAVLT measures. On average, women remembered significantly more words on all primary RAVLT measures. The effect of age was equal across all RAVLT trials. However, the effect of sex and education varied in magnitude across trials. Our results indicated that the female advantage on RAVLT was due to women initially recalling more words on Trial 1 and successfully recalled more words on the subsequent learning trials as reflected by LOT. There was no significant sex-difference on the derived measure LTPR.

In the normative group ($n = 244$), women on average remembered 9.72 words on Trial 7 ($SD = 2.86$) and men remembered 8.15 ($SD = 3.15$). The mean difference of -1.57, 95% CI [-2.33, -0.81] was significant and a two-tailed, independent samples t-test indicated that this difference was significant $t(242) = -4.05, p = <.001$, Cohen's $d = -0.52$. Thus, the female advantage on Trial 7 was estimated to 0.52 SD .

For the derived measure Trials 1-5 total, higher age, lower education, and male sex predicted lower scores (partial $r^2 = .048, p = <.001$, partial $r^2 = .165, p = <.001$, partial $r^2 = .116, p = <.001$, respectively).

For the derived measure LTPR, higher age (partial $r^2 = .029, p = .008$) and lower education (partial $r^2 = .041, p = <.001$) significantly predicted lower scores. For the derived measure LOT, lower education (partial $r^2 = .052, p = <.001$) and male sex (partial $r^2 = .25, p = .014$) predicted lower scores. On the derived measure learning rate (LR), lower education (partial $r^2 = .056, p = <.001$) and male sex predicted lower scores (partial $r^2 = .017, p = .04$). Lastly, proactive and retroactive inhibition were not significantly related to sex, age, or education.

Comparison with published norms from Stricker et al. (2021) in an independent comparison group.

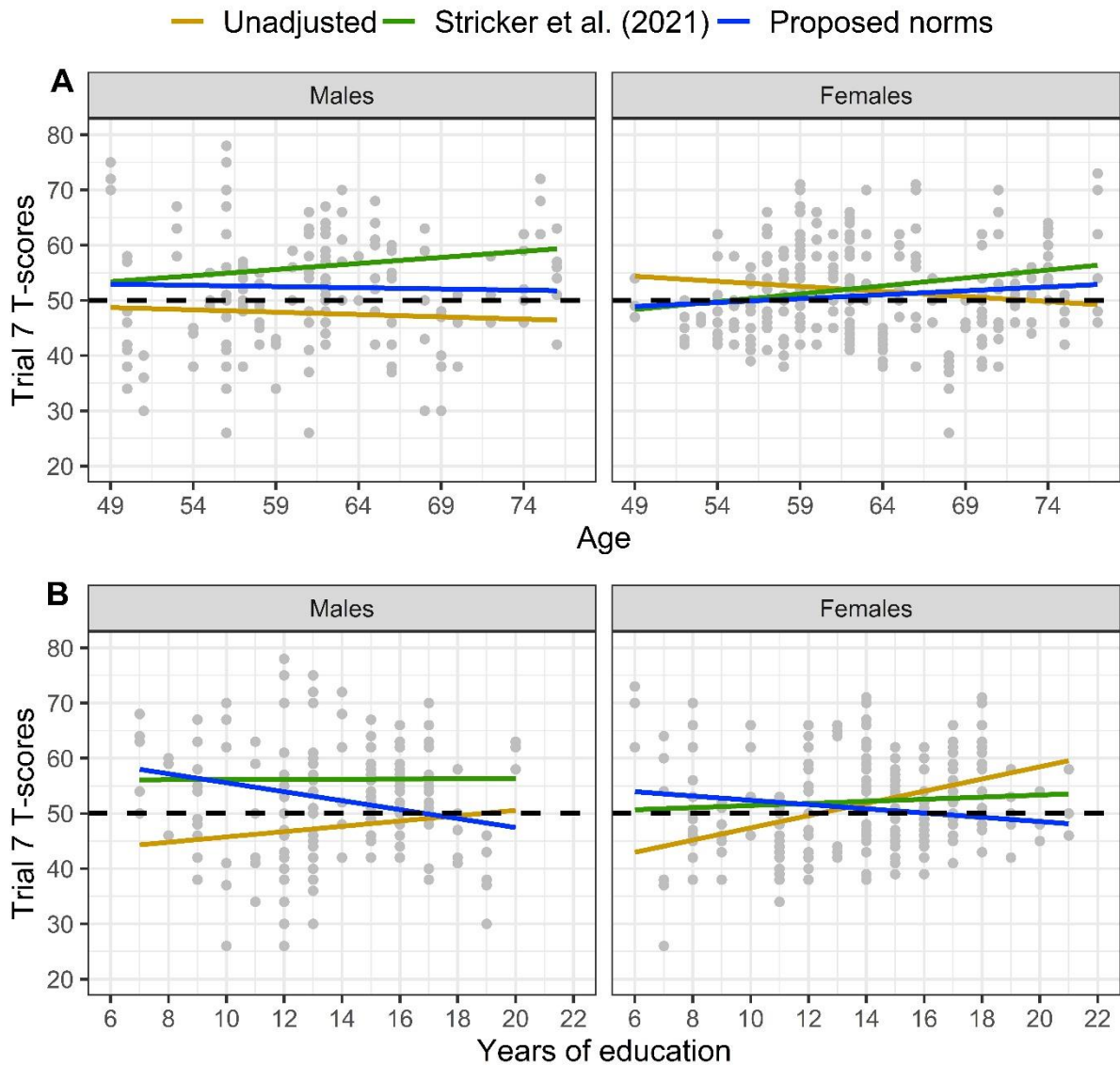
We compared norms from Stricker et al. (2021) and Scandinavian norms in an independent comparison group of cognitively health participants reporting subjective cognitive decline (SCD) ($n = 145$). We compared norms on all primary RAVLT measures and Trials 1-5 total. Regression analyses indicated that norms from Stricker et al. (2021) did not fully adjust for the effects of demographical variables on Trials 1-5 total, Trial 7, Trial 4, or List B. Higher age predicted higher T -scores on Trial 1-5 total (partial $r^2 = .055, p = .005$), Trial 7 (partial $r^2 = .051, p = .007$), Trial 4 (partial $r^2 = .031, p = .037$) and List B (partial $r^2 = .062, p = .003$). Education had no significant linear association with T -scores on any measure. On average, males were estimated 6.9, 4.6, 4.2 and 3.9 higher T -scores than women on Trial 4 (partial $r^2 = .10, p = <.001$), List B (partial $r^2 = .037, p = .022$), Trial 7 (partial $r^2 = .066, p = .002$), and Trials 1-5 total (partial $r^2 = .031, p = .035$), respectively. In comparison, the Scandinavian norms successfully adjusted for the effect of demographics on all trials. Granted, there were linear trends associated with education on Trial 7, however omnibus ANOVAs were not significant ($p >.01$). When using our own norms, males in the sample were on average estimated 0.9, 2.4, 1.6, 2.5 higher T -scores compared to women, but these differences were not significant ($p = >.05$).

In the independent comparison group ($n = 145$) women on average remembered 10.46 words on Trial 7 ($SD = 2.28$) and men remembered 9.41 ($SD = 2.68$). The mean difference of -1.07 95% CI [-1.89, -0.24] was significant and a two-tailed, independent samples t-test indicated that this difference was significant $t(143) = -2.54, p = .012$, Cohen's $d = -0.44$. In other words, the female advantage was estimated to 0.44 SD .

ICCs indicated differing test-retest reliability for basic and derived RAVLT measures in the test-retest sample ($n = 98$). ICCs varied from poor to moderate-good based. The average test-retest interval was 2.5 years ($SD = 0.53$).

Figure 2

Linear plots of RAVLT Trial 7 T-scores computed with Stricker et al (2021) norms, unadjusted scores, and local norms



Note. Independent comparison group of cognitively healthy Scandinavian participants reporting subjective cognitive decline ($n = 145$). Regression lines fitted for age (A) and years of education (B).

5.3 Summary results Paper 3

Effect of demographics on basic CWIT measures.

Scores on CWIT-1, 2, 3, and 4 were significantly related to linear and quadratic effects of age. Here, we observed an accelerated lowering of performance on all CWIT subtests with older age. On CWIT 1-4, partial r^2 values indicated that age and age² explained 9.8%, 2.2%, 21.7%, and 18.4% of the variance in scores, respectively.

On average, women obtained 0.83 higher scaled scores on CWIT-1 (partial $r^2 = .019$, $p = <.001$). On CWIT-3, women obtained 0.45 higher scaled scores on average (partial $r^2 = .007$, $p = .009$).

Years of education was positively related to higher scores on CWIT-3 (partial $r^2 = .008$, $p = .006$) and CWIT-4 (partial $r^2 = .012$, $p = <.001$).

Errors on CWIT-3 and CWIT-4 were not significantly related to age, education, or sex.

Comparison with the original age-adjusted norms from D-KEFS.

The original age-adjusted norms from Delis et al. (2001) did not fully adjust for the effects of age on any of the CWIT subtests in the Norwegian normative sample ($n = 1011$). That is, there were linear effects of age still apparent in the T -scores (Figure 3). On average, higher age predicted higher T -scores on CWIT-1 (partial $r^2 = .008$, $p = .004$), CWIT-2 (partial $r^2 = .032$, $p = <.001$), CWIT-3 (partial $r^2 = .008$, $p = .005$), and CWIT-4 (partial $r^2 = .011$, $p = <.001$). And as expected, the age-adjusted norms from Delis et al. (2001) did not account for the slight female advantage in the Norwegian sample on CWIT-1 (partial $r^2 = .020$, $p = <.001$) and CWIT-3 (partial $r^2 = .006$, $p = .012$). Similarly, the age-adjusted norms failed to adjust for the slight advantage participants with higher educational attainment had on CWIT-3 (partial $r^2 = .008$, $p = .004$) and CWIT-4 (partial $r^2 = .017$, $p = <.001$). On average, the D-KEFS norms estimated T -scores that were slightly above the expected average value of $T = 50$ on CWIT-2, 3, and 4 in the Norwegian sample.

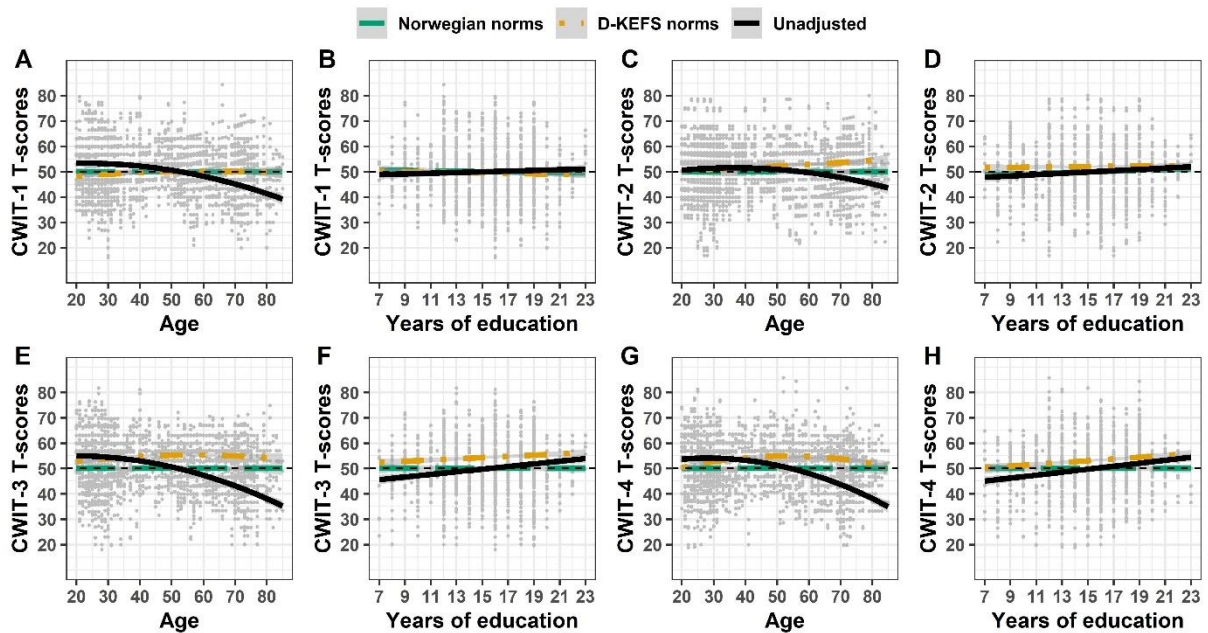
In Paper 3 we compared how many participants in the Norwegian normative sample were deemed as having a score 1.5 SD or more below the normative mean using either our Norwegian norms or the original age-adjusted norms from Delis et al. (2001) (Figure 4). Results indicated that the original age-adjusted norms identified significantly fewer participants compared to the expected base-rate of 6.7% on all CWIT subtests ($p <.01$). In comparison, the Norwegian norms successfully located more participants with scores 1.5 SD

below the normative mean and as expected did not significantly differ compared to the base-rate of 6.7% ($p > .01$). Estimates significantly differed between norms ($p < .001$).

ICCs indicated moderate to good association between baseline and follow-up scores based on an average test-retest interval of 3.4 years ($SD = 0.9$) using the Norwegian norms and the published norms from the US. The test-retest sample comprised 335 participants with available assessments (Table 1).

Figure 3

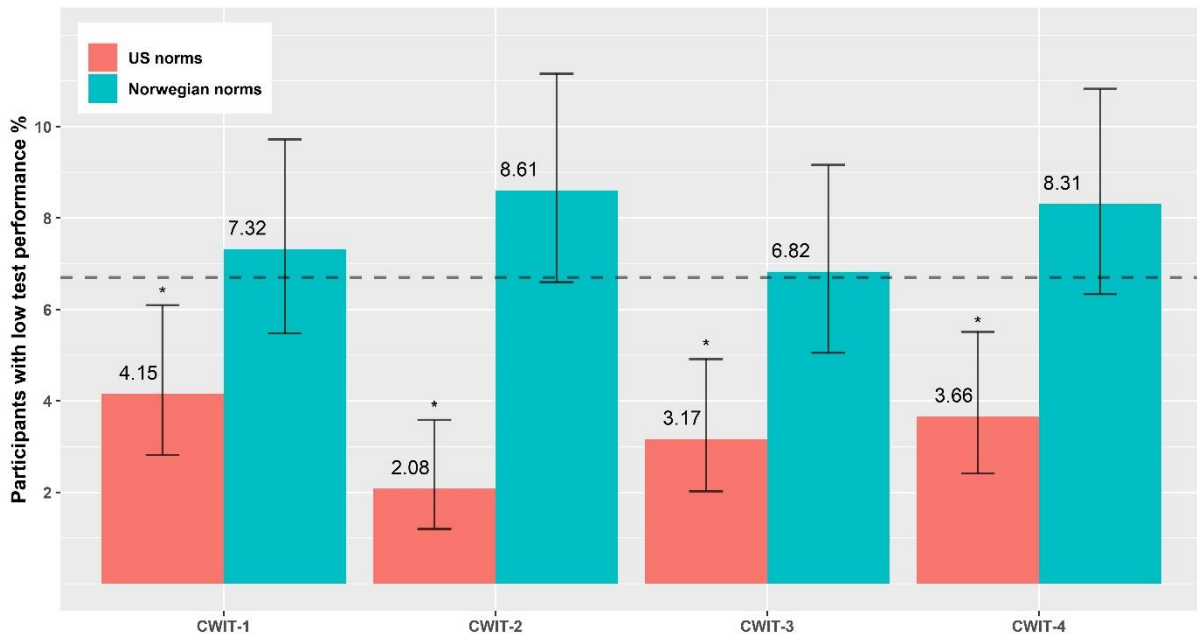
Plots of T-scores on CWIT-1 and CWIT-3 calculated applying norms from Delis et al. (2001), local norms, and T-scores unadjusted for demographic variables



Note. Linear regression lines are fitted for years of education and squared lines for age; for all figures a horizontal line from $T = 50$ represents the ideal normative correction and deviation from this line may indicate maladjustment in the norms.

Figure 4

Percentage of participants in the Norwegian sample ($n = 1011$) with a score 1.5 SD below the normative mean (T -score < 35) on CWIT 1-4



Note. Dotted line indicates the expected base rate for 1.5 SD below the normative mean (6.7%). Error bars indicate the 99% confidence interval (CI) around the estimate. *CI does not contain the expected base rate ($p < .01$). Paired samples proportion tests indicated significant difference between rates from US norms and Norwegian norms on all CWIT subtests ($p < .001$).

6. Discussion

6.1 Summary of findings

Due to a lack of norms or weaknesses associated with earlier norms based on Scandinavian samples, the primary aim of this thesis was a rather practical one; to provide norms on three popular tests frequently used by clinicians and researchers, namely the Trail Making Test (TMT), Rey Auditory Verbal Learning Test (RAVLT) and Color-Word Interference Test (CWIT). We modelled pertinent effects of age, education, and sex on basic and derived measures in Papers 1-3. Statistical assumptions were considered adequate for all normed measures. Our results indicated that all basic scores on TMT, RAVLT, and CWIT were negatively associated with age. That is, we observed a lowering of performance with older age. Education was positively associated with scores on pertinent subtests in line with previous research. Sex differences were observed for several measures relating to verbal

episodic memory on the RAVLT, and executive tasks on the CWIT that specifically involved color-stimuli. To facilitate the use of the regression-based norms for clinicians and researchers we provided norm calculators for every normed measure from Papers 1-3. We hope these new norms will contribute to improved assessment of adult patients and participants in Norway and Sweden.

The secondary aim was to compare the new local norms from Scandinavia with frequently used norms from North America in samples composed of Scandinavians and assess whether the North American norms is suitable in Scandinavian samples and adequately attenuate for demographical variables. In Papers 1-3 we often observed significant linear effects of age, education, and sex on demographically adjusted *T*-scores that were calculated using published norms from North America. In sum, results were indicative of over-adjustment for age (i.e., over-penalizing older participants) and over-adjustment for education (i.e., exaggeration of the difference in test performance between individuals with low and high educational attainment). Regarding sex, we observed less difference between men and women than was expected according to US norms from Stricker et al. (2021) on the RAVLT (Paper 2). In Paper 3, the original age-adjusted norms from Delis et al. (2001) failed to adjust for the slight female advantage observed on CWIT-1 and CWIT-3. Pair-wise comparisons of *T*-scores estimated using either local Scandinavian norms or North American norms in Papers 1-3 indicated significant differences. As a result of the inadequate adjustment of demographical variables, these differences were on average greater for participants with high age, or for participants with either low or high educational attainment. Together the results highlight some consequences of using published norms from North American in these Scandinavian samples. Discrepancies between norms may have clinical implications for the accurate assessments of MCI and cognitive deficits in Scandinavian samples.

6.2 Pattern of age effects in the Scandinavian samples in Papers 1-3

In alignment with most published studies, almost all measures on TMT, CWIT, and RAVLT were negatively influenced by age, i.e., on average there were decreasing scores with higher age. The pattern of age-influence on the measures reported in Papers 1-3 generally align with the theoretical knowledge on the differential sensitivity of cognitive functions for age-related differences (Salthouse, 2010). Broadly speaking, measures relating to fluid intelligence and attention demanding tasks typically deteriorate more with age (Lezak et al., 2012). The age-effect on tests is often conceptualized as a general slowing of information processing capabilities (Salthouse, 1993). This hypothesis, implicating a common factor driving age-

related cognitive decline, is partly supported by the fact that fluid abilities tend to be highly correlated and decrease in conjunction during senescence (Lövdén et al., 2020). As it were, both TMT and CWIT are timed tests with high attentional- and manipulating demands (Berg, Swan, Banks, & Miller, 2016; Sánchez-Cubillo et al., 2009), which may in part explain why these tests were more influenced by age than the RAVLT.

6.2.1 Do published norms from the US underestimate older participants from Scandinavia?

In Papers 1-3 we investigated if published norms from North America successfully adjusted for demographical variables in our Scandinavian samples by regressing *T*-scores calculated using either our local norms or the previously published norms from North America on demographic predictors. The rationale behind these analyses was to assess whether we could detect any patterns of inadequate attenuation in norms that varied according to values of age, sex, or education. Significant linear effects may indicate mis-attenuation for specific ranges of the included predictors. This is not the only criterion of whether norms are suitable. For example, the average scores may consistently differ from the expected mean yet results from *T*-score analysis could indicate no linear effect of demographic variables. As such, this may be considered complementary to the assessment of *T*-score distributions and impairment rates which we will discuss later in section 6.5.

In Paper 1, results indicated that North American norms from Heaton (2004) and Tombaugh (2004) both successfully adjusted for the effects of age. Here, we did not observe any increasing or decreasing mis-attenuation in *T*-scores associated with the age of participants. In Paper 2 (RAVLT) and Paper 3 (CWIT) however, increased age predicted increased *T*-scores calculated with norms from Stricker et al. (2021) and Delis et al. (2001), respectively. The significant linear trend is visualized in Figure 2a, Figure 3a, 3c, 3e and 3g. On the RAVLT and the CWIT, it is apparent from the figures that the older participants were estimated too high *T*-scores relative to the expected normative mean of $T = 50$. In essence, this indicates that the older participants performed better than expected compared to the normative means in Stricker et al. (2021) and Delis et al. (2001). This may indicate that the norms over-adjust for age when applied in our Scandinavian samples. In Paper 2 we used an independent comparison group composed of cognitively healthy participants reporting SCD. We did not observe any linear effect of age when applying our norms on the primary measure Trial 7 (figure 2a). However, when applying the norms from Stricker et al. (2021), age had a significant effect of 5.3% still apparent in the *T*-scores. In comparison, age explained merely 4.8% of the variance in scores in the original normative sample of Scandinavians in Paper 2.

In Paper 3, age explained 0.8%, 3.2%, 0.8%, and 1.1% of the variance in *T*-scores calculated with Delis et al. (2001) norms on CWIT 1-4, respectively. Compared to the total amount of variation in scores associated with age in the Norwegian normative sample on CWIT 1-4 (9.8%, 2.2%, 21.7%, 18.4%, respectively), the in-adequate attenuation by using the Delis et al. (2001) norms were relatively modest.

6.2.2 Why might age-effects be lower in the Scandinavian samples?

In summary, older participants in our Scandinavian samples performed better than expected according to the North American norms in Papers 2, and to a lesser extent in Paper 3. A pertinent question is whether this could represent a generalizable effect characterizing Scandinavian norms, or if it could be explained by other factors related to methodology. To understand why the effect of age might vary between studies we will first consider what adjusting for age on neuropsychological tests represent. Age-effects in norms are cross-sectional in nature, which is not identical to simply adjusting for age-related *decline*. Age-related decline is described in longitudinal studies based on change over time for the same individual. Longitudinal age-related decline (i.e., within-person changes) are typically lower and has less power than cross-sectional age-related *differences* (i.e., between-person differences) (Lövdén et al., 2020). Consider for instance the difference between a group of 70-year-olds and a group of 30-year-olds. This difference may, for instance, entail cohort-effects due to differences in upbringing and time of assessment, educational quality, culture, sensory impairments, motor deficits, in addition to differences in health and medical status. As such, age differences from cross-sectional designs may be more susceptible to variation than one might initially assume. In this regard, it is not unlikely that cohort-differences due to broad cultural and political influences shape the association between age and scores on neuropsychological measures.

So why might the age-effect be lower in the Scandinavian samples in Paper 2 and Paper 3? Firstly, age-effect on scores is often attenuated for when you adjust for cardiovascular disease and other forms of sub-clinical pathology (Harrington et al., 2018; Yu et al., 2015).

Putatively, Norway and the Scandinavian countries have good health care that is available for most, a generally good health status in the public, and high living standards. Indeed, the World Health Organization has for several years listed Norway and Sweden high in terms of Human Development Index (HDI) which is a composite of health, education, and economy (Human Development Reports, 2023). In comparison, health care in the US is not universal and albeit available for many, it is putatively related to the economic status of the individual

(Schneider et al., 2021). We therefore hypothesize that the lower age-difference in the Scandinavian samples could be due to reduced prevalence of vascular disease or other health related factors, potentially due to cultural differences in the availability of health care (Schneider et al., 2021) or other cultural factors influencing the health status of participants.

It is difficult to conclude whether this is a generalizable effect due to cultural differences, or just sample characteristics due to method-biases caused by non-equivalent samples (van de Vijver & Tanzer, 2004). The strongest mis-attenuation with age was observed in Paper 2 when using Stricker et al. (2021) norms on the RAVLT in our independent comparison group. Differences in the screening procedures may have reduced the comparability of samples, thus precluding generalized inferences. Stricker et al. (2021) reports that all participants in the US normative sample were deemed cognitively unimpaired by physician or study coordinator, and participants were recruited from a population register. These assessments were aided by a mental status exam and CDR, however, no cut-offs are specified. While this strategy almost certainly excluded all participants with dementia, this design will likely include participants with varying degrees of cognitive deficits that is present in a normal population. In contrast, all participants in the independent comparison group in Paper 2 were included based on neuropsychological criteria, and no participants in this group scored lower than $T = 35$ on the neuropsychological screening battery because the diagnostic algorithm would otherwise have classified these cases as MCI. As a result, the comparability of the samples is limited due to the differing screening procedures causing the independent comparison group to have higher scores than the normative sample in Stricker et al. (2021).

In Paper 3, the significant age-effect in T -scores calculated using the Delis et al. (2001) norms might be explained by differences due to the time of assessment. The original D-KEFS normative sample were tested between 1998 and 2000. It is well known that the cognitive performance of cohorts generally improve over time, and adults and elderly today perform better on tests than previous generations due to improvements in health and education among others (i.e., Flynn Effect) (Hessel et al., 2018; Skirbekk, Stonawski, Bonsang, & Staudinger, 2013). At the same time, results indicate stagnation or even regression for young adults in western countries (Bratsberg & Rogeberg, 2018). This could be a factor explaining why adults and elderly from Norway performed better than expected according to the Delis et al. (2001) norms. Furthermore, it is possible that Delis et al. (2001) unknowingly included participants with MCI as this was not widely applied as a diagnostic criterion at the time (Petersen, 2004; Petersen et al., 2009; Petersen et al., 1999; Stricker et al., 2021).

While there may be differences in the effect of age in North American norms and Scandinavian norms due to general cohort- and culturally bound differences, it seems unlikely that the tendency for lower age effect compared to the North American norms observed in Paper 2 and Paper 3 could represent a general cultural effect *unique* to Scandinavian participants. For one, we did not observe any faulty adjustments of age compared to Heaton (2004) norms on TMT, although we observed a tendency towards higher-than-expected scores on TMT-B using Heaton (2004) norms (Figure 1c). Also, compared to Tombaugh (2004) norms, we observed no significant effect of age on *T*-scores, although the norms were generally too strict for Scandinavians of all ages on TMT-B (Figure 1c). Furthermore, the effect of age in the Scandinavian samples on TMT, RAVLT or CWIT was not substantially lower compared to published studies outside North America. In our sample of participants between 41 and 84 years we found on TMT-A and TMT-B that age explained 17% and 15.9% of the variance in scores, respectively. Recent normative studies in comparable age-ranges frequently report effect sizes around this range (García-Herranz et al., 2022; Llinàs-Reglà et al., 2017; Lojo-Seoane et al., 2023) or lower (St-Hilaire et al., 2018). On the RAVLT, our results indicated an age-effect of 4.8% on Trials 1-5 total in our normative sample between 49 – 79 years. Again, this is comparable to several international studies (Bezdicek et al., 2014; Dassanayake et al., 2020) although Stricker et al. (2021) and others find greater age effect on scores (Ferreira Correia & Campagna Osorio, 2014; Knight, McMahon, Green, & Skeaff, 2006; Messinis et al., 2016). Lastly, in Paper 3 on the CWIT, studies using Stroop tests have found less effect of age in comparable age-ranges (Bezdicek et al., 2015; Dassanayake et al., 2021; Magnúsdóttir et al., 2021) but most find higher effect in the same age-range (Ktaiche et al., 2022; Rodríguez-Lorenzana et al., 2021; Wim Van der Elst et al., 2006; Vicente et al., 2021).

6.3 Effect of education in the Scandinavian samples in Papers 1-3 and comparison with North American norms

Results from multiple regression analyses on *T*-scores showed linear trends indicating inadequate attenuation that varied according to the years of education of participants in Paper 1 using Heaton (2004) norms on TMT-A and TMT-B. On average, higher education predicted lower *T*-scores on TMT-A (partial $r^2 = 7.9\%$) and TMT-B (partial $r^2 = 6.8\%$). From Figure 1b and 1d, it is apparent that participants with low educational attainment received too high *T*-scores relative to the expected mean, and reversely participants with high educational attainment received too low *T*-scores. This indicates that the Heaton (2004) norms expected

participants with low educational attainment to perform worse, and participants with high educational attainment to perform better. This over-adjustment is likely because education was more closely associated with scores in the initial normative sample of Heaton (2004). Results from our normative sample indicated no significant effect of education on TMT-A, and only a small effect of 3.6% explained variance on TMT-B. In contrast, Heaton (2004) reported 10% on TMT-A and 16% on TMT-B.

The Tombaugh (2004) norms for the TMT were adjusted for education in two levels (0-12 and >12 years) exclusively for participants over the age of sixty. This stratification is rather crude, but we saw no linear effect in which years of education predicted *T*-scores using Tombaugh (2004) norms in the Scandinavian normative sample. Tombaugh (2004) reports that there was very little effect of education after controlling for age for the younger participants (<55 years) and therefore only adjust for education for participants over 60 years of age where the effects of education reportedly were more pronounced (1.5% and 4.4% on TMT-A and TMT-B, respectively). Likely, it is because the educational effect was so low in the Scandinavian sample that this crude adjustment for education provided an adequate attenuation.

In Paper 3 on the CWIT the effects associated with education were also low in the normative sample. Results indicated no significant effect of education on CWIT-1 or CWIT-2, and only 0.8% and 1.2% variance on CWIT-3 and 4, respectively. The age-adjusted norms from Delis et al. (2001) were not adjusted for education, and we observed slight mis-attenuation relating to education in the *T*-scores in the normative sample on CWIT-3 (partial $r^2 = 0.8\%$) and CWIT-4 (partial $r^2 = 1.7\%$). Again, this is likely because the effect of education in the Norwegian sample was so weak.

In Paper 2 our results indicated that education had a large effect compared to age and sex on the summary measure Trials 1-5 total (16.5% variance) and observed a pattern in which delayed memory indices (Trial 7) were more associated with education than the initial learning trials (Trials 1 through 5). The Stricker et al. (2021) norms were adjusted for education, and our results indicated no linear effect associated with education in the *T*-scores in the Scandinavian independent comparison group. Compared to age and sex, the effect of education on scores in the Stricker et al. (2021) sample was relatively weak. Although, the way we coded education for Scandinavian participants was not precisely aligned with the procedures in the Stricker et al. (2021) norms which could have influenced these results. For instance, participants with a bachelor's degree in Norway would in most cases be coded as

having either 15- or 16-years total education (12/13 years basic schooling + 3 years bachelor's degree). In contrast, Stricker et al. (2021) specifies that a bachelor's degree is to be coded as either 16 or 17 years depending on if any additional graduate schooling was attained by the participants. As a result, using years of education as a continuous variable, the Scandinavian participants may have been compared to American participants with a lower educational *level*. However, in most cases the coding of education was in alignment, and it is doubtful whether this had any profound effect on scores due to the relatively weak effect associated with education in the Stricker et al. (2021) norms.

Compared to recently published normative studies outside North America on the TMT, RAVLT, and CWIT, the educational effects in our Scandinavian samples are indeed lower than commonly reported by most studies. Results from Paper 1 indicated no significant effects associated with education TMT-A and 3.6% explained variance on TMT-B. This is much lower than previously reported by most (García-Herranz et al., 2022; Llinàs-Reglà et al., 2017; Lojo-Seoane et al., 2023) but not all studies (St-Hilaire et al., 2018).

On the RAVLT, the effect associated with education was estimated to 16.5% explained variance on the summary measure Trials 1-5 total. This is mostly comparable albeit slightly higher than other studies (Bezdicek et al., 2014; Dassanayake et al., 2020; Ferreira Correia & Campagna Osorio, 2014; Messinis et al., 2016).

Lastly, on the CWIT there were no norms outside the original age-adjusted norms from Delis et al. (2001), but recent studies on comparable Stroop paradigms indicate much higher influence of education in norms than apparent in the Norwegian sample in Paper 3 (Dassanayake et al., 2020; Ktaiche et al., 2022; Magnúsdóttir et al., 2021; Rodríguez-Lorenzana et al., 2021; Wim Van der Elst et al., 2006; Vicente et al., 2021).

6.3.1 Are Scandinavian norms characterized by lower influence of education?

In sum, it appears that norms which are either not adjusted for education, or just weakly adjusted for education, provided an adequate fit in the Scandinavians samples in Papers 1-3. Furthermore, compared to other published normative studies the effects of education are generally lower than expected. Similar results were obtained by Kirsebom et al. (2019) in a partially overlapping sample from Paper 1 and 2 on the CERAD delayed recall test. While our results and these comparisons with previously published norms are not suitable for conclusive inferences about broad cultural differences, we hypothesize that this could represent a trend characterizing norms in Scandinavia.

To explain why education might have weaker associations with neuropsychological scores in the Scandinavian samples we must consider how education is thought to relate to cognitive test scores. First, education has consistent correlations with IQ (Ritchie & Tucker-Drob, 2018; Steinberg, Bieliauskas, Smith, & Ivnik, 2005) and according to a review by Lövdén et al. (2020) adolescents with an initial high IQ tend to have high educational attainment later in life. Secondly, a meta-analysis assessing the effect of several quasi-experimental studies on the effects of policy changes increasing state mandated education have shown that one additional year of education during adolescence was associated with an average increase of 2 IQ points later in adulthood (Ritchie & Tucker-Drob, 2018). Furthermore, education might be causally related to neurobiological adaptations early in life that pervades through adulthood (Lövdén et al., 2020) as cross-sectional associations indicate that increased educational attainment is related to increased brain reserve (i.e., higher cortical volume) (Nyberg et al., 2021; Stern et al., 2023). Thirdly, the positive effects of education on cognitive tests are established early in life and are then upheld and remain stable throughout adulthood via: increased access to cognitively engaging occupations; socioeconomic advantages; increased access to health care leading to lower mortality; and decreased dementia risk via brain reserve (Ceci & Williams, 1997; Fratiglioni & Wang, 2007; Montez, Hummer, & Hayward, 2012). In consideration of these putative mechanisms, it is likely that cultural factors related to the accessibility, quality, average length, and secular benefits gained from education throughout adulthood are moderated by cultural and political factors. Indeed, Lövdén et al. (2020) notes that the psychological correlates of between-person differences in educational attainment may vary with age, cohort, period, and society. For one, the tendency for adolescents with higher IQ to seek higher education might be moderated by the fact that in Norway and Scandinavian countries, university level education is financially supported by the state, and many studies are open without any minimum grade requirements (Samordna opptak, 2023). Secondly, due to the organization of the welfare state in Scandinavian countries and low pay-gaps, the secular benefits from having a high educational attainment might be lower in Scandinavian countries compared to many others (Statistics Norway, 2023b). In sum, we hypothesize that education could be a proxy for different factors in Norway and other Scandinavian countries compared to international studies due to an egalitarian education policy where socioeconomic status is not heavily dictated by educational attainment, scholastic aptitude may not be critical for accessing higher education, in conjunction with an in-discriminatory health care service that is available for most regardless of socioeconomic status (Schneider et al., 2021). In turn, this moderates the relationship between education and cognitive test scores and causes

education to appear as having a lower effect in Scandinavia. More studies are needed to confirm or deny the tendency observed in Papers 1-3, preferably with representative samples using harmonized procedures.

6.4 Sex differences in Scandinavian samples in Paper 2

In Paper 2, results indicated a considerable sex difference in favor of women on RAVLT. On average, women remembered 1.47 more words on Trial 7, adjusted for differences in age and education. Our results indicate that women remembered more words on Trial 1, reflecting better attentional ability (Woodard, 2006, pp. 105–142), and amassed more words over the subsequent learning trials (reflected by LOT). However, women did not differ significantly from men in their ability to remember previously learned material after 30 minutes as there were no significant differences in the measure LTPR. Our results are in accordance with another Norwegian study showing that the female advantage in verbal episodic memory is mediated by improved auditory attention span and inhibitory control of irrelevant stimuli, but not short-term memory per se (Kljajevic et al., 2023).

A female advantage in verbal episodic memory is consistently reported (Weber et al., 2014), however the magnitude is known to vary between countries (Asperholm et al., 2019). In a large meta regression analysis based on 495 studies from 45 countries, Asperholm et al. (2019) found positive univariate associations between greater female advantage in verbal episodic memory and increased gender equality, increased gross domestic product (GDP) per capita, and increased population education and employment rates. In these analyses, Norway and Sweden scored among the highest on these metrics and had corresponding large sex differences in verbal episodic memory. In fact, out of all 45 considered countries, the difference between men and women was greatest in Norway. This result was based on 12 Norwegian studies and the average effect size was estimated to approximately 0.6 *SD*, and in Sweden the effect size was estimated to about 0.35 *SD* based on 22 studies. In our own normative sample of Norwegians and Swedes, the combined difference between men and women was estimated to 0.52 *SD* which corresponds well to the meta-analytic estimates in Asperholm et al. (2019).

6.4.1 Adjustment of sex-differences using Stricker et al. (2021) norms

Assessment of *T*-scores estimated using US norms from Stricker et al. (2021) indicated significant sex-differences in the independent comparison group on Trials 1-5 total, Trial 4, Trial 7, and list B. The Stricker et al. (2021) norms were considerably stricter for women. Women were on average estimated between 0.4 – 0.7 *SD* lower *T*-scores on the above-

mentioned subtests. As described previously, all participants in independent comparison group in Paper 2 obtained scores above $T = 35$ on an abbreviated neuropsychological test battery. We therefore expected slightly elevated scores in the independent comparison group, but the sex-difference in T -scores should be adjusted for regardless of the overall mean in the sample. In comparison, using our own norms, women obtained on average between 0.1 – 0.3 SD lower T -scores than men, even though this sex-difference was not related to significant omnibus ANOVAs. Thus, it appears that our Scandinavian norms were better able to adjust for the sex-difference in the independent comparison group. The reason why our own norms still produced some sex-difference in T -scores was because the magnitude of the sex-difference in the independent comparison group was smaller (0.44 SD) than in the normative group (0.52 SD). In the Scandinavian norms we therefore expected a greater difference between men and women than was apparent in the independent comparison group. As a result, women ended up obtaining slightly lower T -scores than men using our Scandinavian norms and the Stricker et al. (2021) norms. This is indicative of an overadjustment in the norms when applied in the independent comparison group.

The Stricker et al. (2021) norms do not report separate raw scores for men and women but Figure 1 in Stricker et al. (2021) indicate that the difference between men and women were greater in the US. In the Stricker et al. (2021) norms, women remembered approximately 2 words more than men on Trial 7 which is more than in the Scandinavian normative group and the independent comparison group. As a result, the Stricker et al. (2021) norms over-adjusted for the difference between men and women in the independent comparison group. The mis-attenuation by Stricker et al. (2021) is unlikely due to differences in age and education between the US normative group and our Scandinavian sample, as the coefficients from multiple regression analysis describe the differences between men and women adjusted for differences in age or education and no significant interactions were reported. It is surprising that Stricker et al. (2021) reports a stronger sex effect considering the aforementioned meta-analysis by Asperholm et al. (2019) which indicated that the sex-differences were larger in Norway and Sweden compared to the US. In the meta-analysis by Asperholm et al. (2019) the average sex-difference in the US was estimated to about 0.25 SD . The Stricker et al. (2021) study is a very large population-based study ($n = 4428$), so it is unlikely that the sex-difference observed in their sample is a spurious result. It is not apparent from the Asperholm et al. (2019) study how big the variance surrounding the meta-analytic estimate was. The difference between Stricker et al. (2021) and the meta-analytic estimate could be due to high

variance. Interestingly, previous studies in Norway on comparable episodic memory tests have repeatedly shown that the sex-difference is weaker in Norway than in the US, causing women to have lower mean *T*-scores than men (Egeland et al., 2005; Kanestrøm, 2017). Thus, our results and the results from previous Norwegian investigations do not align with the results from the Asperholm et al. (2019) meta-analysis. Regardless of the reason why, our results indicate that sex-differences in our Scandinavian normative sample was weaker than in the US norms by Stricker et al. (2021) which highlights the importance of using local norms.

6.5 Clinical implications and suitability of the North American norms in Scandinavian samples

In Paper 1 we observed that the Tombaugh (2004) norms were generally too strict compared to the expected mean on TMT-B in the Scandinavian sample (average difference approx. -0.5 *SD*). This was influenced by some participants that were estimated very poor scores ($< T$ 20) in the Tombaugh (2004) norms that skewed distributions. As a result of the traditional norming methodology, certain normed cells of age and education in the Tombaugh (2004) norms had very low variance, and any departure from the cell mean caused these very poor scores ($< T$ 20). Traditional norming methodologies as employed here are prone to produce misleading results (Van Breukelen & Vlaeyen, 2005). The overall clinical implication is that we would on average expect more false positive MCI cases using these norms in Scandinavian samples. Using the Heaton (2004) norms, the largest discrepancy between norms were observed for individuals in the end ranges of predictors. In other words, participants with high or low educational attainment. The clinical implications are that norms will on average be less sensitive for assessing MCI in individuals with low education (i.e., produce more false negatives) and will cause too many false positives for individuals with high education. This is concerning because TMT is one of the most frequently used tests internationally (Kreutzer et al., 2011; Rabin et al., 2005) and in Norway (Vaskinn & Egeland, 2012).

In Paper 2, older participants from Scandinavia performed better than expected on RAVLT Trial 7 when applying norms from Stricker et al. (2021). This is concerning because delayed recall measures like Trial 7 on the RAVLT is considered one of the best measures for assessing verbal episodic memory deficits in amnesic MCI due to AD (Estévez-González, Kulisevsky, Boltes, Otermín, & García-Sánchez, 2003; Landau et al., 2010; Vuoksimaa et al., 2020). Furthermore, it may be especially important that assessments are valid for participants over the age of 60 because episodic memory decline due to AD typically manifest from this

age onwards (Bassett & Folstein, 1993) and age is the strongest predictor for dementia (Morris, Clark, & Vissel, 2018). Thus, the mis-attenuation from Stricker et al. (2021) norms might have important clinical implications for accurate diagnosis of amnesic MCI in adults and elderly in Scandinavia. Furthermore, results from Paper 2 indicated an inadequate attenuation for sex using Stricker et al. (2021) norms (Figure 2). We therefore expect that in Scandinavian samples, using these norms will on average result in too many false negatives for men and too many false positives for women.

In Paper 3 we assessed the proportion of participants obtaining a low score (defined as a score 1.5 *SD* below the normative mean) using the original age adjusted norms from Delis et al. (2001) and our Norwegian norms. Result indicated significant differences between the expected theoretical base rate (~6.7%) and the estimated proportion of participants with low scores when applying the Delis et al. (2001) norms. Compared to our Norwegian norms, the Delis et al. (2001) norms located significantly fewer participants with low scores. This indicates that the Delis et al. (2001) norms might have a lower sensitivity for accurately detecting cognitive deficits in the Norwegian sample which might have direct implications for the accurate assessment of patients.

Lastly, an overarching clinical implication from Papers 1-3 is that Norwegian and Swedish psychologists and other users of norms can make comparisons with improved cultural appropriateness. In fact, since the year 2000 the Norwegian Psychologists Association has mandated that members use tests which are technically sound and appropriate for the situation with norms that are representative for the target group (Commission, 2001; Ryder, 2021). This is an important implication from an ethical standpoint as no investigations that we are aware of had previously assessed how these published norms performed in Scandinavian countries despite previous studies showing significant differences between local norms and other published norms (Egeland et al., 2005; Fernandez & Marcopulos, 2008; Lorentzen et al., 2023; Raudeberg et al., 2019). Results from Paper 1-3 indicate significant differences between published norms and local norms in accordance with the many studies which stress the importance of local norms in neuropsychological assessment (Hayden et al., 2014; Hestad et al., 2016; Ojeda et al., 2016; Rivera et al., 2015; Weber et al., 2014). With the norms presented in Papers 1-3, Norwegian and Swedish users of norms can be more certain on the cultural validity of measures when norm-referencing patients in their assessments.

6.6 Improvements on traditional derived measures by employing regression-based approaches

In Paper 1 we proposed a novel way of calculating the derived measure TMT B-A which we called TMT- β . This measure can be considered a further development of previous norms by Senior, Piovesana, and Beaumont (2018) who suggested an improvement on the conventional TMT B-A measure in what they called ‘stratified discrepancy scores’. The basic premise was that in the presence of pathology (i.e., in clinical samples), TMT-A completion times and TMT-B completion times are non-linearly related as TMT-B completion times increase more due to increased distance between stimuli and more potential distractions (Gaudino et al., 1995). Therefore, in clinical samples with slow completion times on TMT-A and TMT-B, the result is an apparent large TMT B-A score. Because norms on TMT B-A are made based on healthy participants where the relationship between TMT-A and TMT-B is linearly related, Senior, Piovesana, and Beaumont (2018) argued that applying normed TMT B-A scores in clinical samples erroneously give the impression of executive deficits. The authors provided traditional norms stratified for age in two levels (≥ 50 years, <50 years), education in two levels (≥ 12 years, <12 years), and TMT-A performance in three levels (fast, average, slow). The authors demonstrated that in a large heterogeneous clinical sample, not adjusting for performance on TMT-A when assessing TMT B-A scores led to misrepresentation of executive dysfunction. They found that a common outcome occurring for 37% of participants in the clinical sample was that both TMT-A and TMT-B scores were slow relative to healthy participants. For this subset, analyzing the conventional TMT B-A score resulted in an erroneously conclusion there was executive dysfunction in 16% of all cases in the clinical sample. However, analyzing TMT B-A stratified for TMT-A completion times allowed Senior, Piovesana, and Beaumont (2018) to conclude that 40% of these did not in fact have difficulties with the additional task demands of TMT-B, but primarily had difficulty with speed of visual search and perceptual speed. This is because their completion time on TMT-B was not abnormally long compared to others with similar TMT-A completion times.

As demonstrated by Senior, Piovesana, and Beaumont (2018), it is clearly beneficial to be able to discern whether elevated TMT B-A scores is due to deficits in higher order cognitive functions, visual search and perceptual speed, or both. However, the norms proposed by Senior, Piovesana, and Beaumont (2018) suffer from inherent disadvantages due to the traditional norming methodology (Van Breukelen & Vlaeyen, 2005). Firstly, the Senior, Piovesana, and Beaumont (2018) norms had small cell sizes for some combinations of

predictors. In fact, multiple cells had very small sample sizes (e.g., $n = 3$, $n = 12$, $n = 13$) due to the unusual combination of predictors. This is not optimal because the normative statistics are computed directly from these discrete distributions. Furthermore, the arbitrary stratification of continuous variables lowers the precision of demographic adjustments and may result in the edge of cohort effect (Crompvoets et al., 2021).

In contrast, in Paper 1 we proposed norms on TMT-B adjusting pertinent demographical variables and TMT-A scores in a continuous fashion using multiple regression analysis. We called this measure TMT- β to differentiate from the conventional methods and highlight the regression-based approach. This should simultaneously resolve the issues with the conventional TMT B-A approach and further improve on the traditional stratified approach described by Senior, Piovesana, and Beumont (2018). TMT B-A and TMT- β were strongly correlated in our normative sample, which is assuring considering the useful properties of TMT B-A (Devora, Beevers, Kiselica, & Bengel, 2019). The lack of differentiation between TMT- β and TMT B-A may be explained by the results from Senior, Piovesana, and Beaumont (2018) which show that such approaches are mainly beneficial in clinical samples. After Paper 1 was published in 2020, others have also implemented a continuous approach like TMT- β with apparent benefits (Iñesta, Oltra-Cucarella, Bonete-López, Calderón-Rubio, & Sitges-Maciá, 2021).

Similarly, because the D-KEFS test battery is primarily conceptualized as a tool for assessing executive functions the first two subtests of the CWIT are often considered control tests or baseline conditions (D. C. Delis et al., 2001; Halleland et al., 2012). Thus, the primary interest is often on inhibition (CWIT-3) and inhibition/switching (CWIT-4) (Lezak et al., 2012). However, performance on CWIT-3 and CWIT-4 relies upon the basic abilities in the first task. One study using the D-KEFS battery showed that most of the apparent difference in patients' executive function deficits (i.e., differences on CWIT-3 and CWIT-4) could be explained by differences in lower-order functions (i.e., differences on CWIT-1 and CWIT-2) (Savla et al., 2011). Comparing performance on CWIT-1 or CWIT-2 with performance on CWIT-3 and CWIT-4 may be used in a process-oriented interpretation to isolate the higher-order functions and assess whether the deficient performance is mainly due to basic abilities or higher-order cognitive abilities. D. C. Delis et al. (2001) included norms for such contrast scores in the D-KEFS manual. However, these measures are reported to have poor reliability due to the computation method and it is not advised using these measures in clinical assessments (Crawford, Sutherland, & Garthwaite, 2008; Lezak et al., 2012). The issue with the Delis et

al. (2001) contrast scores is that the two measures involved in the subtraction each have their own unique measurement error, but the contrast score absorbs both of these measurement errors additively (Crawford et al., 2008). However, it is possible to calculate these measures in a regression-based approach similar to TMT- β by analyzing scores on CWIT-3 or CWIT-4 regressed on CWIT-1 and CWIT-2, and pertinent demographical variables (Steinunn Adólfssdóttir et al., 2014; Halleland et al., 2012) which might alleviate some of the issues with the original contrast measures proposed by Delis et al. (2001). In fact, this was suggested in an earlier publication by Wim Van der Elst et al. (2006) and we are currently constructing norms for such measures based on a Norwegian sample (Lorentzen et al., in prep.). This way of assessing scores on CWIT-3 and CWIT-4 has shown clinical utility for assessing executive deficits in patient with ADHD (Halleland et al., 2012). Furthermore, such measures have shown differential sensitivity to the measurement of cortical volumes in healthy adults (Pa et al., 2010). Steinunn Adólfssdóttir et al. (2014) reported significant correlations between task performance on adjusted CWIT-3 and CWIT-4 scores (but not basic CWIT-3 and CWIT-4 scores) and cortical volume in the middle frontal gyrus (MFG) and dorsolateral prefrontal cortex (DLPFC) in healthy Norwegian adults. These are areas frequently implicated in executive processes (Krueger et al., 2011; McDonald, Delis, Norman, Tecoma, & Iragui-Madoz, 2005).

To summarize, regression-based approaches to isolate task demands shows promise and the utility of such measures are supported in previous studies. Furthermore, regression-based approaches may alleviate issues with reliability associated with conventional subtraction methods such as TMT B-A and contrast measures by Delis et al. (2001). As far as we are aware, Paper 1 presents the first norms on TMT-B adjusting for performance on TMT-A in a continuous fashion. We are not aware of any similar norms on the CWIT either. Moving forward there is a need to assess the validity and reliability of these measures in local samples and assess the incremental utility of these above the standard measures.

6.7 Test-retest reliability in Paper 2 and Paper 3

Typically, test-retest reliability estimates are conceptualized as estimates of how independent a score is of measurement error, assuming no true change in the underlying construct has occurred in between measurements (Polit, 2014). However, in Paper 2 and Paper 3 we calculated test-retest reliability estimates based on lengthy test intervals. Estimates from long intervals may be more aptly characterized as measures of temporal stability (A. Evers et al., 2013) or stability over time as we did in Paper 3. As a result, our local estimates are likely a

mixture of both true change and random measurement error. Under these conditions, we would naturally expect lower reliability estimates than other studies with shorter test intervals (Polit, 2014).

For the most important basic measures in Paper 2, namely RAVLT Trial 7 and Trials 1-5 total, estimates were ‘moderate’ to ‘good’ in relation to regular criterion (Koo & Li, 2016). The individual learning trials (Trial 1 through Trial 5) had lower reliability, indicating that they are more susceptible to random error variance. This is typical for measures with high attentional demands which may be considered ‘changeable’ and prone to variation (Sherman, Brooks, Iverson, Slick, & Strauss, 2010; Woodard, 2006). Clinical decisions based on these measures in isolation is therefore cautioned. Most of the derived measures on the RAVLT had inadequate reliability and caution is advised when interpreting these measures. In relation to test evaluation guidelines the reliability of the most important measures were ‘adequate’ (A. Evers et al., 2013). However, the sample size in the test-retest group was marginally lower than the accepted range ($n = 98$ vs. $n = 100$). In Paper 3, ICCs from normed scores on CWIT 1-4 were all moderate to good and considered adequate based on test evaluation guidelines even with the prolonged test-retest interval ($M = 3.4$ years, $SD = 0.9$ years) (A. Evers et al., 2013; Koo & Li, 2016). We hope that providing evidence of reliability on these measures in local samples can bolster the confidence of Norwegian and Swedish test users of the psychometric properties of the norms.

6.8 Methodical considerations and study limitations in Papers 1-3

The current thesis is not without limitations. We took great care in assessing model assumptions and modelling pertinent effects of demographic variables in the normative samples and therefore believe our norms adequately describe the relationship between scores on the cognitive measures and demographic variables in the normative samples. However, the external validity, i.e., how well the norms transfer to other sample populations in Norway and Sweden is an important consideration. In Papers 1-3 participants were predominantly recruited through advertisements in newspapers, senior organizations, symptom group participants, and an orthopedic ward. As such, the samples may be considered self-selected convenience samples. This is quite common in normative research, however there might be unknown biases associated with non-probability convenience sampling (Jager, Putnick, & Bornstein, 2017). Anecdotally, many of the healthy participants agreed to participate out of altruistic motives, frequently because their spouse or other relative had suffered from dementia, and not necessarily because they expected to perform well on the

neuropsychological tests. This may limit some of the self-selection bias in relation to the cognitive performance of the samples. The various eligibility criteria for participation also affect the external validity of the norms. For instance, all participants reported either Norwegian or Swedish as native language. Ethnicity was not recorded; however, almost all participants were of European ethnicity. As a result, we expect norms to be less accurate for participants with native languages other than Norwegian or Swedish, and individuals of foreign ethnicity such as immigrants or refugees, even though they constitute a substantial proportion of the Norwegian society. In 2023, it is estimated that almost six hundred thousand immigrants from countries outside the Nordics live in Norway which equates to ~11% of the adult population (Statistics Norway, 2023a).

Furthermore, there are considerations regarding the cognitive status of participants. We did not formally assess depressive symptoms using symptoms scales with defined cut-offs, however major depression was an exclusion criterion in Papers 1-3. As a result, there might be participants with depressive symptoms above standardized cut-off scores that were included. In Papers 1 and 2, MMSE was used to include participants (≥ 26). In Paper 3, MMSE was not used as a stringent cut-off, but aided in the evaluation of a participant as healthy and almost all participants had high scores on the MMSE ($M = 29.1$, $SD = 1.1$, range = 24 – 30). Out of the 1011 participants in Paper 3, 40 had missing values on the MMSE. It seems unlikely that this had a large influence as the substantial sample size in conjunction with assessment of outliers and influential cases limit the influence of individual observations (Kwak & Kim, 2017). Nevertheless, we did not have longitudinal records to confirm the cognitive normalcy of participants over time. Based on the inclusion/exclusion criteria, our samples in Paper 1-3 sit somewhere in between what might be considered undeniably healthy, so-called robust norms (Bos et al., 2018; Sliwinski, Lipton, Buschke, & Stewart, 1996) and pure population-based sample (i.e., unscreened). Both have advantages and disadvantages as we expect a heavily screened sample to have over-all better diagnostic accuracy (but lower sensitivity) (Bos et al., 2018), and a pure population-based sample to have good sensitivity but low specificity (O'Connell & Tuokko, 2010). Another limitation is that we did not have an independent comparison group in Paper 1 and 3 to assess the norms in. Yet, we have some indication that Norwegian participants in Paper 1 and Paper 2 performed better than age-matched participants from a Norwegian population-based study on the CERAD recall test (Wagle et al., 2023).

In sum, it is important for users of the norms to be aware that the norms naturally do not reflect all adults in the age range, but that the eligibility criteria and recruitment methods set the boundaries for the sample population and by extension the generalizability. In our opinion, the norms in Paper 1-3 are probably representative of many of the adults and elderly that are referred to neuropsychological assessments. The normative samples in Paper 1-3 may perform better or worse than other sample populations in Norway or Sweden. As a result, it is important that clinicians and researchers use norms from samples that resemble the intended population and make informed decisions on the appropriateness of norms (Heaton, Avitable, Grant, & Matthews, 1999).

In Paper 3 the normative sample had higher educational attainment than expected in the public ($M = 15.5$, $SD = 2.9$). According to Statistics Norway the educational attainment in the population is distributed into three approximately equal parts; mandatory schooling (<10 years); high school and trade schools (<13 years); and University degrees (>14 years) (Espenes et al., 2022). This is not necessarily detrimental to the representativeness of the norms as effects of lower educational attainment were modeled by including individuals with lower levels of educational attainment as well (education range = 7-23 years). This is because regression-based norming is not significantly affected by unbalanced datasets like traditional norming methodologies (Wim Van der Elst et al., 2006). In multiple regression analysis, imbalance does not bias the unstandardized beta coefficients, but instead increases the standard errors causing reduced power and increased p -values (Kleinbaum et al., 2013). A limitation in this regard is that we did not construct confidence intervals to accurately display the uncertainty of normed scores (H. E. M. Oosterhuis, van der Ark, & Sijtsma, 2017; Lieke Voncken, Casper J. Albers, & Marieke E. Timmerman, 2019).

Furthermore, a potential limitation is that we did not assess whether including education as a categorical predictor changed the interpretation of scores. Even though education is often adjusted for as a continuous variable in norms, this makes theoretical assumptions on the linear relationship between education and test scores that may not be justifiable in all samples. For instance, it is possible that the effect of education on scores is diminishing for higher educational levels (Hankee et al., 2016; Lezak et al., 2012). However, model fits using education as a continuous variable were overall good. Furthermore, we did not observe any polynomial effects of education which might suggest a diminishing association between neuropsychological scores and increasing educational attainment. Another limitation concerning the statistical analyses is that we did not perform any multiple-test adjustments to

reduce the family-wise error rate such as Bonferroni-type adjustments or adjusting for false discovery rate (Glickman, Rao, & Schultz, 2014; James et al., 2013). For the normative analyses, the risk of our approach is of course false positive results and thus norming measures according to irrelevant variables. To combat false positive results, we predominately reported exact p -values for the normative analyses in Paper 1-3 and applied a more stringent criterion of $\alpha = .01$ in Paper 3. In many cases, the associated p -values were much smaller ($p < .001$). As recommended, p -values always analyzed in conjunction with previous results, theoretical expectations, and in consideration of the associated effect sizes (Feise, 2002). Also, in case of many statistical analyses, as in Paper 2 when comparing norms from Stricker et al. (2021) and our local norms, we interpreted p -values from omnibus ANOVAs and not individual coefficients to reduce the risk of chance capitalization.

Another limitation concerns the comparison of local norms and US norms regarding the adequate attenuation of demographic variables in Papers 1-3. In these analyses we did not investigate whether coefficients from multiple regression analysis on the T -scores calculated using US norms and local norms significantly differed from each other. For instance, while the effect of education was still apparent in T -scores applying Heaton (2004) norms, but not when applying our local norms, this does not imply that the effect of education on scores significantly differed in the analyses (Gelman & Stern, 2006). Although in most cases the difference in coefficients was substantial and we believe it is likely that they would reach threshold for statistical significance had equivalence of the coefficients been tested.

Lastly, no participants reported significant sensory, or motor impairments and testing did not proceed if participants had sensory impairments that hindered completion of tasks.

Participants were instructed to wear hearing aids and glasses whenever pertinent. However, this was not formally assessed, and we cannot guarantee that this did not influence the normative estimates. For instance, as might be expected, visual acuity is known to significantly disturb performance on the TMT and it is recommended to thoroughly assess the visual acuity of participants and/or patients (Fröhlich, Müller, & Voelcker-Rehage, 2021).

6.9 Future directions

The results from this thesis show consistent differences between international and local norms which may indicate a need to update norms on other neuropsychological tests frequently in use as well (Ryder, 2021). Furthermore, other populations in Norway may require specific norms such as groups of immigrants and refugees, which often are not validly assessed using norms based on the ethnic majority (Franzen et al., 2022). The issues with incompatibility of

published norms in Norway raises the question of whether development of norms and assessing validity and reliability based on these norms should be up to clinicians, researchers, and test publishers, or if the relevant health authorities should have a more active role in the assessment and development of norms to ensure good quality neuropsychological assessment. Secondly, while we present norms in Papers 1-3, we did not have an independent comparison group to assess norms in, and the independent group in Paper 2 was sourced from the same cohort. Preferably, future studies should assess the norms in population-based samples from Norway or other Scandinavian countries. Such studies may also allow for broader inferences about cultural differences characterizing Norwegian and Scandinavian norms. The validity of the norms in Paper 1-3 needs to be assessed in relevant samples based on test evaluation guidelines (Commission, 2001; A. Evers et al., 2013). Furthermore, the incremental utility and validity of TMT- β should be assessed in clinical samples in accordance with test evaluation guidelines and previous studies on similar measures (Steinunn Adólfssdóttir et al., 2014; Brewster, Pasqualini, & Martin, 2022; Halleland et al., 2012; Pa et al., 2010; Senior et al., 2018).

7. Conclusions

In this thesis we presented norms on basic and derived indices of TMT, RAVLT and D-KEFS CWIT based on samples of Norwegian and Swedish healthy adults. Test scores were significantly related to age, education, and sex in general alignment with previous studies. In Papers 1-3 we provided clinicians and researchers with normative calculators adjusting for pertinent effects of demographical variables. Compared to North American norms, the effects of age and education were typically weaker in the Scandinavian samples. On the RAVLT, differences between men and women were lower in the Scandinavian samples compared to published norms from the US. Due to incompatibility of samples and method biases in normative research we are unable to conclude whether this represents broad cultural differences characterizing Scandinavian norms. Our results highlight some implications of using North American norms in Scandinavian samples and indicate a need for further development of local norms on neuropsychological measures for Scandinavian populations.

8. References

- Adólfssdóttir, S., Haász, J., Wehling, E., Ystad, M., Lundervold, A., & Lundervold, A. J. (2014). Salient measures of inhibition and switching are associated with frontal lobe gray matter volume in healthy middle-aged and older adults. *Neuropsychology*, *28*, 859-869. doi:10.1037/neu0000082
- Adólfssdóttir, S., Wollschlaeger, D., Wehling, E., & Lundervold, A. J. (2017). Inhibition and Switching in Healthy Aging: A Longitudinal Study. *J Int Neuropsychol Soc*, *23*(1), 90-97. doi:10.1017/s1355617716000898
- Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., . . . Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*, *7*(3), 270-279. doi:10.1016/j.jalz.2011.03.008
- Alviarez-Schulze, V., Cattaneo, G., Pachón-García, C., Solana-Sánchez, J., Tormos, J. M., Pascual-Leone, A., & Bartrés-Faz, D. (2022). Validation and Normative Data of the Spanish Version of the Rey Auditory Verbal Learning Test and Associated Long-Term Forgetting Measures in Middle-Aged Adults. *Front Aging Neurosci*, *14*, 809019. doi:10.3389/fnagi.2022.809019
- Arbuthnott, K., & Frank, J. (2000). Executive control in set switching: residual switch cost and task-set inhibition. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *54*(1), 33.
- Asperholm, M., Nagar, S., Dekhtyar, S., & Herlitz, A. (2019). The magnitude of sex differences in verbal episodic memory increases with social progress: Data from 54 countries across 40 years. *PLoS One*, *14*(4), e0214945.
- Barbosa, R., Midão, L., Almada, M., & Costa, E. (2021). Cognitive performance in older adults across Europe based on the SHARE database. *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn*, *28*(4), 584-599. doi:10.1080/13825585.2020.1799927
- Bassett, S. S., & Folstein, M. F. (1993). Memory complaint, memory performance, and psychiatric diagnosis: a community study. *Journal of geriatric psychiatry and neurology*, *6*(2), 105-111.
- Battery, A. I. T. (1944). Manual of directions and scoring. In: Washington, DC: War Department, Adjutant General's Office.
- Baxendale, S. (2010). The Flynn effect and memory function. *Journal of Clinical and Experimental Neuropsychology*, *32*(7), 699-703. Retrieved from <https://doi.org/10.1080/13803390903493515>. doi:10.1080/13803390903493515
- Beck, A. T., Steer, R. A., & Brown, G. K. (1987). *Beck depression inventory*: Harcourt Brace Jovanovich New York:.
- Beier, M., Hughes, A. J., Williams, M. W., & Gromisch, E. S. (2019). Brief and cost-effective tool for assessing verbal learning in multiple sclerosis: Comparison of the Rey Auditory Verbal Learning Test (RAVLT) to the California Verbal Learning Test - II (CVLT-II). *J Neurol Sci*, *400*, 104-109. doi:10.1016/j.jns.2019.03.016
- Berg, J.-L., Swan, N. M., Banks, S. J., & Miller, J. B. (2016). Atypical performance patterns on Delis-Kaplan Executive Functioning System Color-Word Interference Test: Cognitive switching and learning ability in older adults. *Journal of Clinical and Experimental Neuropsychology*, *38*(7), 745-751. Retrieved from <https://doi.org/10.1080/13803395.2016.1161734>. doi:10.1080/13803395.2016.1161734
- Bezdicek, O., Lukavsky, J., Stepankova, H., Nikolai, T., Axelrod, B. N., Michalec, J., . . . Kopecek, M. (2015). The Prague Stroop Test: Normative standards in older Czech adults and discriminative validity for mild cognitive impairment in Parkinson's

- disease. *J Clin Exp Neuropsychol*, 37(8), 794-807.
doi:10.1080/13803395.2015.1057106
- Bezdicek, O., Motak, L., Axelrod, B. N., Preiss, M., Nikolai, T., Vyhnalek, M., . . . Ruzicka, E. (2012). Czech Version of the Trail Making Test: Normative Data and Clinical Utility. *Archives of Clinical Neuropsychology*, 27(8), 906-914. Retrieved from <https://doi.org/10.1093/arclin/acs084>. doi:10.1093/arclin/acs084
- Bezdicek, O., Stepankova, H., Axelrod, B. N., Nikolai, T., Sulc, Z., Jech, R., . . . Kopecek, M. (2017). Clinimetric validity of the Trail Making Test Czech version in Parkinson's disease and normative data for older adults. *Clin Neuropsychol*, 31(sup1), 42-60. doi:10.1080/13854046.2017.1324045
- Bezdicek, O., Stepankova, H., Moták, L., Axelrod, B. N., Woodard, J. L., Preiss, M., . . . Poreh, A. (2014). Czech version of Rey Auditory Verbal Learning test: normative data. *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn*, 21(6), 693-721. doi:10.1080/13825585.2013.865699
- Biundo, R., Weis, L., Pilleri, M., Facchini, S., Formento-Dojot, P., Vallelunga, A., & Antonini, A. (2013). Diagnostic and screening power of neuropsychological testing in detecting mild cognitive impairment in Parkinson's disease. *J Neural Transm (Vienna)*, 120(4), 627-633. doi:10.1007/s00702-013-1004-2
- Boake, C. (2000). Edouard Claparede and the auditory verbal learning test. *Journal of Clinical and Experimental Neuropsychology*, 22(2), 286-292.
- Boenniger, M. M., Staerk, C., Coors, A., Huijbers, W., Ettinger, U., & Breteler, M. M. B. (2021). Ten German versions of Rey's auditory verbal learning test: Age and sex effects in 4,000 adults of the Rhineland Study. *J Clin Exp Neuropsychol*, 43(6), 637-653. doi:10.1080/13803395.2021.1984398
- Bondi, M. W., Edmonds, E. C., Jak, A. J., Clark, L. R., Delano-Wood, L., McDonald, C. R., . . . for the Alzheimer's Disease Neuroimaging, I. (2014). Neuropsychological Criteria for Mild Cognitive Impairment Improves Diagnostic Precision, Biomarker Associations, and Progression Rates. *Journal of Alzheimer's Disease*, 42, 275-289. doi:10.3233/JAD-140276
- Bondi, M. W., Serody, A. B., Chan, A. S., Ebersone-Shumate, S. C., Delis, D. C., Hansen, L. A., & Salmon, D. P. (2002). Cognitive and neuropathologic correlates of Stroop Color-Word Test performance in Alzheimer's disease. *Neuropsychology*, 16, 335-343. doi:10.1037/0894-4105.16.3.335
- Bos, I., Vos, S. J. B., Jansen, W. J., Vandenberghe, R., Gabel, S., Estanga, A., . . . Visser, P. J. (2018). Amyloid- β , Tau, and Cognition in Cognitively Normal Older Individuals: Examining the Necessity to Adjust for Biomarker Status in Normative Data. *Front Aging Neurosci*, 10, 193. doi:10.3389/fnagi.2018.00193
- Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. *Proc Natl Acad Sci U S A*, 115(26), 6674-6678. doi:10.1073/pnas.1718793115
- Brewster, B. M., Pasqualini, M. S., & Martin, L. E. (2022). Functional Brain Connectivity and Inhibitory Control in Older Adults: A Preliminary Study. *Front Aging Neurosci*, 14, 763494. doi:10.3389/fnagi.2022.763494
- Cavaco, S., Gonçalves, A., Pinto, C., Almeida, E., Gomes, F., Moreira, I., . . . Teixeira-Pinto, A. (2013). Trail Making Test: Regression-based norms for the Portuguese population. *Archives of Clinical Neuropsychology*, 28(2), 189-198.
- Ceci, S. J., & Williams, W. M. (1997). Schooling, intelligence, and income. *American Psychologist*, 52(10), 1051-1058. doi:10.1037/0003-066X.52.10.1051
- Cerezo García, M., Martín Plasencia, P., & Aladro Benito, Y. (2015). Alteration profile of executive functions in multiple sclerosis. *Acta Neurol Scand*, 131(5), 313-320. doi:10.1111/ane.12345

- Clark, L. R., Schiehser, D. M., Weissberger, G. H., Salmon, D. P., Delis, D. C., & Bondi, M. W. (2012). Specific Measures of Executive Function Predict Cognitive Decline in Older Adults. *Journal of the International Neuropsychological Society*, *18*(1), 118-127. Retrieved from <https://www.cambridge.org/core/article/specific-measures-of-executive-function-predict-cognitive-decline-in-older-adults/C40844BCFEAA9FFFDE5BDC2316C9007F>. doi:10.1017/S1355617711001524
- Commission, I. T. (2001). International guidelines for test use. *International Journal of testing*, *1*(2), 93-114.
- Crawford, J. R., Sutherland, D., & Garthwaite, P. H. (2008). On the reliability and standard errors of measurement of contrast measures from the D-KEFS. *Journal of the International Neuropsychological Society*, *14*(6), 1069-1073.
- Crompvoets, E. A. V., Keuning, J., & Emons, W. H. M. (2021). Bias and Precision of Continuous Norms Obtained Using Quantile Regression. *Assessment*, *28*(6), 1735-1750. doi:10.1177/1073191120910201
- Dassanayake, T. L., Hewawasam, C., Baminiwatta, A., & Ariyasinghe, D. I. (2021). Regression-based, demographically adjusted norms for victoria stroop test, digit span, and verbal fluency for Sri Lankan adults. *The Clinical Neuropsychologist*, *35*(sup1), S32-S49.
- Dassanayake, T. L., Hewawasam, C., Baminiwatta, A., Samarasekara, N., & Ariyasinghe, D. I. (2020). Sex-, age-and education-adjusted norms for the WHO/UCLA version of the Rey Auditory Verbal Learning Test for Sinhala-speaking Sri Lankan adults. *The Clinical Neuropsychologist*, *34*(sup1), 127-142.
- de Lange, A. G., Bråthen, A. C. S., Rohani, D. A., Fjell, A. M., & Walhovd, K. B. (2018). The Temporal Dynamics of Brain Plasticity in Aging. *Cereb Cortex*, *28*(5), 1857-1865. doi:10.1093/cercor/bhy003
- Delis, D. (2005). Delis-Kaplan executive function system, Norwegian version. *Bromma: Pearson Assessment*.
- Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan executive function system*.
- Derogatis, L. R., & Unger, R. (2010). Symptom checklist-90-revised. *The Corsini encyclopedia of psychology*, 1-2.
- Devora, P. V., Beevers, S., Kiselica, A. M., & Benge, J. F. (2019). Normative Data for Derived Measures and Discrepancy Scores for the Uniform Data Set 3.0 Neuropsychological Battery. *Archives of Clinical Neuropsychology*, *35*(1), 75-89. Retrieved from <https://doi.org/10.1093/arclin/acz025>. doi:10.1093/arclin/acz025
- Dickinson, M. D., & Hiscock, M. (2011). The Flynn effect in neuropsychological assessment. *Appl Neuropsychol*, *18*(2), 136-142. doi:10.1080/09084282.2010.547785
- Duchek, J. M., Balota, D. A., Thomas, J. B., Snyder, A. Z., Rich, P., Benzinger, T. L., . . . Ances, B. M. (2013). Relationship between Stroop performance and resting state functional connectivity in cognitively normal older adults. *Neuropsychology*, *27*(5), 516.
- Eckerström, C., Olsson, E., Bjerke, M., Malmgren, H., Edman, Å., Wallin, A., & Nordlund, A. (2013). A combination of neuropsychological, neuroimaging, and cerebrospinal fluid markers predicts conversion from mild cognitive impairment to dementia. *Journal of Alzheimer's Disease*, *36*(3), 421-431.
- Egeland, J., Sundet, K., Landrø, N. I., Rund, B. R., Asbjørnsen, A., Hugdahl, K., . . . Stordal, K. (2005). Validering av normer for oversatte tester av oppmerksomhet og hukommelse i et norsk normalutvalg. *Tidsskrift for Norsk psykologforening*, *42*(2).
- Eliassen, I. V., Fladby, T., Kirsebom, B.-E., Eckerström, M., Wallin, A., Bråthen, G., . . . Hessen, E. (2020). Predictive and diagnostic utility of brief neuropsychological

- assessment in detecting Alzheimer's pathology and progression to dementia. *Neuropsychology*, 34(8), 851.
- Engedal, K., Benth, J. S., GjØra, L., Skjellegrind, H. K., Nāvik, M., & Selbæk, G. (2023). Normative Scores on the Norwegian Version of the Mini-Mental State Examination. *Journal of Alzheimer's Disease*(Preprint), 1-12.
- Espenes, J., Eliassen, I. V., Øhman, F., Hessen, E., Waterloo, K., Eckerstrøm, M., . . . Kirsebom, B.-E. (2022). Regression-based normative data for the Rey Auditory Verbal Learning Test in Norwegian and Swedish adults aged 49–79 and comparison with published norms. *The Clinical Neuropsychologist*, 1-25. Retrieved from <https://doi.org/10.1080/13854046.2022.2106890>. doi:10.1080/13854046.2022.2106890
- Espenes, J., Hessen, E., Eliassen, I. V., Waterloo, K., Eckerstrøm, M., Sando, S. B., . . . Kirsebom, B.-E. (2020). Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years. *The Clinical Neuropsychologist*, 34(sup1), 110-126. Retrieved from <https://doi.org/10.1080/13854046.2020.1829068>. doi:10.1080/13854046.2020.1829068
- Espinosa, A., Alegret, M., Valero, S., Vinyes-Junqué, G., Hernández, I., Mauleón, A., . . . Tárraga, L. (2013). A longitudinal follow-up of 550 mild cognitive impairment patients: evidence for large conversion to dementia rates and detection of major risk factors involved. *J Alzheimers Dis*, 34(3), 769-780.
- Estévez-González, A., Kulisevsky, J., Boltes, A., Otermín, P., & García-Sánchez, C. (2003). Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: comparison with mild cognitive impairment and normal aging. *International Journal of Geriatric Psychiatry*, 18(11), 1021-1028. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/gps.1010>. doi:<https://doi.org/10.1002/gps.1010>
- Evers, A. (2012). The Internationalization of Test Reviewing: Trends, Differences, and Results. *International Journal of testing*, 12(2), 136-156. Retrieved from <https://doi.org/10.1080/15305058.2012.658932>. doi:10.1080/15305058.2012.658932
- Evers, A., Muñiz, J., Hagemester, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the Quality of Tests: Revision of the EFPA Review Model. *Psicothema*, 25(3), 283-291. doi:10.7334/psicothema2013.97
- Fábián, B., Kenyhercz, F., Bugán, A., & Andrejkovics, M. (2023). [Normative data on clinical neuropsychological tests in Hungary I.]. *Orv Hetil*, 164(15), 577-585. doi:10.1556/650.2023.32705
- Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology*, 2(1), 8. Retrieved from <https://doi.org/10.1186/1471-2288-2-8>. doi:10.1186/1471-2288-2-8
- Fernandez, A. L., & Marcopulos, B. A. (2008). A comparison of normative data for the Trail Making Test from several countries: Equivalence of norms and considerations for interpretation. *Scandinavian journal of psychology*, 49(3), 239-246.
- Ferreira Correia, A., & Campagna Osorio, I. (2014). The Rey Auditory Verbal Learning Test: normative data developed for the Venezuelan population. *Arch Clin Neuropsychol*, 29(2), 206-215. doi:10.1093/arclin/act070
- Fillenbaum, G. G., van Belle, G., Morris, J. C., Mohs, R. C., Mirra, S. S., Davis, P. C., . . . Welsh-Bohmer, K. A. (2008). Consortium to Establish a Registry for Alzheimer's Disease (CERAD): the first twenty years. *Alzheimer's & dementia*, 4(2), 96-109.
- Fine, E. M., Delis, D. C., & Holdnack, J. (2011). Normative Adjustments to the D-KEFS Trail Making Test: Corrections for Education and Vocabulary Level. *The Clinical*

- Neuropsychologist*, 25(8), 1331-1344. Retrieved from <https://doi.org/10.1080/13854046.2011.609838>. doi:10.1080/13854046.2011.609838
- Fjell, A. M., Chen, C.-H., Sederevicius, D., Sneve, M. H., Grydeland, H., Krogsrud, S. K., . . . Walhovd, K. B. (2018). Continuity and Discontinuity in Human Cortical Development and Change From Embryonic Stages to Old Age. *Cerebral Cortex*, 29(9), 3879-3890. Retrieved from <https://doi.org/10.1093/cercor/bhy266>. doi:10.1093/cercor/bhy266
- Fladby, T., Pålhaugen, L., Selnes, P., Bråthen, G., Hessen, E., Almdahl, I. S., . . . Espenes, R. (2017). Detecting at-risk Alzheimer's disease cases. *Journal of Alzheimer's Disease*, 60(1), 97-105.
- Floden, D., Vallesi, A., & Stuss, D. T. (2011). Task context and frontal lobe activation in the Stroop task. *Journal of Cognitive Neuroscience*, 23(4), 867-879.
- Flynn, J. R. (1987). Massive IQ Gains in 14 Nations: What IQ Tests Really Measure. *Psychological bulletin*, 101(2), 171-191. doi:10.1037/0033-2909.101.2.171
- 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. *J Psychiatr Res*, 12(3), 189-198. doi:10.1016/0022-3956(75)90026-6
- Formanek, T., Kagstrom, A., Winkler, P., & Cermakova, P. (2019). Differences in cognitive performance and cognitive decline across European regions: a population-based prospective cohort study. *Eur Psychiatry*, 58(58), 80-86. doi:10.1016/j.eurpsy.2019.03.001
- Franzen, S., Watermeyer, T. J., Pomati, S., Papma, J. M., Nielsen, T. R., Narme, P., . . . Bekkhus-Wetterberg, P. (2022). Cross-cultural neuropsychological assessment in Europe: Position statement of the European Consortium on Cross-Cultural Neuropsychology (ECCroN). *The Clinical Neuropsychologist*, 36(3), 546-557. Retrieved from <https://doi.org/10.1080/13854046.2021.1981456>. doi:10.1080/13854046.2021.1981456
- Fratiglioni, L., & Wang, H.-X. (2007). Brain reserve hypothesis in dementia. *Journal of Alzheimer's Disease*, 12(1), 11-22.
- Fröhlich, S., Müller, K., & Voelcker-Rehage, C. (2021). Normative Data for the CERAD-NP for Healthy High-Agers (80-84 years) and Effects of Age-Typical Visual Impairment and Hearing Loss. *J Int Neuropsychol Soc*, 1-13. doi:10.1017/s1355617721001284
- Fällman, K., Lundgren, L., Wressle, E., Marcusson, J., & Classon, E. (2020). Normative data for the oldest old: Trail Making Test A, Symbol Digit Modalities Test, Victoria Stroop Test and Parallel Serial Mental Operations. *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn*, 27(4), 567-580. doi:10.1080/13825585.2019.1648747
- García-Herranz, S., Díaz-Mardomingo, M. C., Suárez-Falcón, J. C., Rodríguez-Fernández, R., Peraita, H., & Venero, C. (2022). Normative Data for Verbal Fluency, Trail Making, and Rey–Osterrieth Complex Figure Tests on Monolingual Spanish-Speaking Older Adults. *Archives of Clinical Neuropsychology*, 37(5), 952-969. Retrieved from <https://doi.org/10.1093/arclin/acab094>. doi:10.1093/arclin/acab094
- Gary, S., Lenhard, W., & Lenhard, A. (2021). Modelling norm scores with the cNORM package in R. *Psych*, 3(3), 501-521.
- 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.
- Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, 60(4), 328-331.
- Gjora, L., Strand, B. H., Bergh, S., Borza, T., Braekhus, A., Engedal, K., . . . Selbaek, G. (2021). Current and Future Prevalence Estimates of Mild Cognitive Impairment, Dementia, and Its Subtypes in a Population-Based Sample of People 70 Years and

- Older in Norway: The HUNT Study. *J Alzheimers Dis*, 79(3), 1213-1226.
doi:10.3233/JAD-201275
- Glickman, M. E., Rao, S. R., & Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology*, 67(8), 850-857. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0895435614001127>.
doi:<https://doi.org/10.1016/j.jclinepi.2014.03.012>
- Golden, C., Freshwater, S. M., & Golden, Z. (1978). Stroop color and word test.
- Halleland, H. B., Haavik, J., & Lundervold, A. J. (2012). Set-Shifting in Adults with ADHD. *Journal of the International Neuropsychological Society*, 18(4), 728-737. Retrieved from <https://www.cambridge.org/core/article/setshifting-in-adults-with-adhd/50515B8609A3D3BDE515DD9BBF5C5792>. doi:10.1017/S1355617712000355
- Hankee, L. D., Preis, S. R., Piers, R. J., Beiser, A. S., Devine, S. A., Liu, Y., . . . Au, R. (2016). Population Normative Data for the CERAD Word List and Victoria Stroop Test in Younger- and Middle-Aged Adults: Cross-Sectional Analyses from the Framingham Heart Study. *Exp Aging Res*, 42(4), 315-328.
doi:10.1080/0361073x.2016.1191838
- Harrington, K. D., Schembri, A., Lim, Y. Y., Dang, C., Ames, D., Hassenstab, J., . . . Maruff, P. (2018). Estimates of age-related memory decline are inflated by unrecognized Alzheimer's disease. *Neurobiology of Aging*, 70, 170-179. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0197458018302094>.
doi:<https://doi.org/10.1016/j.neurobiolaging.2018.06.005>
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2): Springer.
- Hayden, K. M., Makeeva, O. A., Newby, L. K., Plassman, B. L., Markova, V. V., Dunham, A., . . . Roses, A. D. (2014). A comparison of neuropsychological performance between US and Russia: preparing for a global clinical trial. *Alzheimers Dement*, 10(6), 760-768.e761. doi:10.1016/j.jalz.2014.02.008
- Heaton, R. K. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults, professional manual*: Psychological Assessment Resources.
- Heaton, R. K., Avitable, N., Grant, I., & 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(4), 572-582.
- Hessel, P., Kinge, J. M., Skirbekk, V., & Staudinger, U. M. (2018). Trends and determinants of the Flynn effect in cognitive functioning among older individuals in 10 European countries. *Journal of Epidemiology and Community Health*, 72(5), 383-389. Retrieved from <https://jech.bmj.com/content/jech/72/5/383.full.pdf>. doi:10.1136/jech-2017-209979
- Hessen, E., Reinvang, I., Eliassen, C. F., Nordlund, A., Gjerstad, L., Fladby, T., & Wallin, A. (2014). The combination of dysexecutive and amnesic deficits strongly predicts conversion to dementia in young mild cognitive impairment patients: A report from the Gothenburg-Oslo MCI Study. *Dementia and geriatric cognitive disorders extra*, 4(1), 76-85.
- Hestad, K. A., Menon, J. A., Serpell, R., Kalungwana, L., Mwaba, S. O. C., Kabuba, N., . . . Heaton, R. K. (2016). Do neuropsychological test norms from African Americans in the United States generalize to a Zambian population? *Psychological Assessment*, 28, 18-38. doi:10.1037/pas0000147

- Hester, R. L., Kinsella, G. J., Ong, B., & McGregor, J. (2005). Demographic influences on baseline and derived scores from the trail making test in healthy older Australian adults. *The Clinical Neuropsychologist*, *19*(1), 45-54.
- Human Development Reports. (2023). Human Development Index (HDI). Retrieved from <https://hdr.undp.org/data-center/human-development-index#/indicies/HDI>
- Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992). Mayo's older Americans normative studies: updated AVLT norms for ages 56 to 97. *The Clinical Neuropsychologist*, *6*(S1), 83-104.
- Jager, J., Putnick, D. L., & Bornstein, M. H. (2017). II. MORE THAN JUST CONVENIENT: THE SCIENTIFIC MERITS OF HOMOGENEOUS CONVENIENCE SAMPLES. *Monogr Soc Res Child Dev*, *82*(2), 13-30. doi:10.1111/mono.12296
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.
- Jessen, F., Amariglio, R. E., Van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., . . . Van Der Flier, W. M. (2014). A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimer's & dementia*, *10*(6), 844-852.
- Kanestrøm, H. (2017). California Verbal Learning Test (CVLT-II) og Brief Visuospatial Memory Test – Revised (BVM-T-R): Undersøkelse av prestasjoner og samsvar i et klinisk utvalg. *Tidsskrift for Norsk Nevropsykologisk Forening*, *19*(1), 10–16.
- Karr, J. E., Garcia-Barrera, M. A., Holdnack, J. A., & Iverson, G. L. (2018). Advanced clinical interpretation of the Delis-Kaplan Executive Function System: multivariate base rates of low scores. *The Clinical Neuropsychologist*, *32*(1), 42-53. Retrieved from <https://doi.org/10.1080/13854046.2017.1334828>. doi:10.1080/13854046.2017.1334828
- Keifer, E., & Tranel, D. (2013). A neuropsychological investigation of the Delis-Kaplan executive function system. *Journal of Clinical and Experimental Neuropsychology*, *35*(10), 1048-1059.
- Kenyhercz, F., Fábrián, B., Andrejkovics, M., & Bugán, A. (2023). [Normative data on clinical neuropsychological tests in Hungary II.]. *Orv Hetil*, *164*(16), 618-629. doi:10.1556/650.2023.32706
- Kinge, J. M., Dieleman, J. L., Karlstad, Ø., Knudsen, A. K., Klitkou, S. T., Hay, S. I., . . . Vollset, S. E. (2023). Disease-specific health spending by age, sex, and type of care in Norway: a national health registry study. *BMC Medicine*, *21*(1), 201. Retrieved from <https://doi.org/10.1186/s12916-023-02896-6>. doi:10.1186/s12916-023-02896-6
- Kirsebom, B.-E., Espenes, R., Hessen, E., Waterloo, K., Johnsen, S. H., Gundersen, E., . . . Fladby, T. (2019). Demographically adjusted CERAD wordlist test norms in a Norwegian sample from 40 to 80 years. *The Clinical Neuropsychologist*, *33*(sup1), 27-39.
- 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*(1), 77-82.
- Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2013). *Applied regression analysis and other multivariable methods*: Cengage Learning.
- Kljajević, V., Evensmoen, H. R., Sokołowski, D., Pani, J., Hansen, T. I., & Håberg, A. K. (2023). Female advantage in verbal learning revisited: a HUNT study. *Memory*, *31*(6), 831-849. Retrieved from <https://doi.org/10.1080/09658211.2023.2203431>. doi:10.1080/09658211.2023.2203431
- Knight, R. G., McMahan, J., Green, T. J., & Skeaff, C. M. (2006). Regression equations for predicting scores of persons over 65 on the Rey Auditory Verbal Learning Test, the

- mini-mental state examination, the trail making test and semantic fluency measures. *Br J Clin Psychol*, 45(3), 393-402. doi:10.1348/014466505X68032
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2), 155-163.
- Kreutzer, J. S., DeLuca, J., & Caplan, B. (2011). *Encyclopedia of clinical neuropsychology*: Springer.
- Krueger, C. E., Laluz, V., Rosen, H. J., Neuhaus, J. M., Miller, B. L., & Kramer, J. H. (2011). Double dissociation in the anatomy of socioemotional disinhibition and executive functioning in dementia. *Neuropsychology*, 25(2), 249.
- Ktaiche, M., Fares, Y., & Abou-Abbas, L. (2022). Stroop color and word test (SCWT): Normative data for the Lebanese adult population. *Appl Neuropsychol Adult*, 29(6), 1578-1586. doi:10.1080/23279095.2021.1901101
- Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology*, 70(4), 407-411.
- Landau, S. M., Harvey, D., Madison, C. M., Reiman, E. M., Foster, N. L., Aisen, P. S., . . . Initiative, O. b. o. t. A. s. D. N. (2010). Comparing predictors of conversion and decline in mild cognitive impairment. *Neurology*, 75(3), 230-238. Retrieved from <https://n.neurology.org/content/neurology/75/3/230.full.pdf>. doi:10.1212/WNL.0b013e3181e8e8b8
- Lavoie, M., Bherer, L., Joubert, S., Gagnon, J. F., Blanchet, S., Rouleau, I., . . . Hudon, C. (2018). Normative data for the Rey Auditory Verbal Learning Test in the older French-Quebec population. *Clin Neuropsychol*, 32(sup1), 15-28. doi:10.1080/13854046.2018.1429670
- Lee, N. R., Wallace, G. L., Raznahan, A., Clasen, L. S., & Giedd, J. N. (2014). Trail making test performance in youth varies as a function of anatomical coupling between the prefrontal cortex and distributed cortical regions. *Front Psychol*, 5, 496. doi:10.3389/fpsyg.2014.00496
- Lenhard, A., Lenhard, W., Suggate, S., & Segerer, R. (2018). A Continuous Solution to the Norming Problem. *Assessment*, 25(1), 112-125. Retrieved from <https://journals.sagepub.com/doi/abs/10.1177/1073191116656437>. doi:10.1177/1073191116656437
- Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment, 5th ed.* New York, NY, US: Oxford University Press.
- Lippa, S. M., & Davis, R. N. (2010). Inhibition/switching is not necessarily harder than inhibition: An analysis of the D-KEFS color-word interference test. *Archives of Clinical Neuropsychology*, 25(2), 146-152.
- Llinàs-Reglà, J., Vilalta-Franch, J., López-Pousa, S., Calvó-Perxas, L., Torrents Rodas, D., & Garre-Olmo, J. (2017). The Trail Making Test: Association with other neuropsychological measures and normative values for adults aged 55 years and older from a Spanish-speaking population-based sample. *Assessment*, 24(2), 183-196.
- Lojo-Seoane, C., Facal, D., Delgado-Losada, M. L., Rubio-Valdehita, S., López-Higes, R., Frades-Payo, B., & Pereiro, A. X. (2023). Normative scores for attentional tests used by the Spanish consortium for ageing normative data (SCAND) study: Trail Making Test, Digit Symbol and Letter Cancellation. *Clin Neuropsychol*, 1-21. doi:10.1080/13854046.2023.2173304
- Lorentzen, I. M., Espenes, J., Hessen, E., Waterloo, K., Bråthen, G., Timón, S., . . . Kirsebom, B.-E. (2023). Regression-based norms for the FAS phonemic fluency test for ages 40–84 based on a Norwegian sample. *Applied Neuropsychology: Adult*, 30(2), 159-168. Retrieved from <https://doi.org/10.1080/23279095.2021.1918128>. doi:10.1080/23279095.2021.1918128

- Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The Importance of the Normality Assumption in Large Public Health Data Sets. *Annual Review of Public Health, 23*(1), 151-169. Retrieved from <https://doi.org/10.1146/annurev.publhealth.23.100901.140546>. doi:10.1146/annurev.publhealth.23.100901.140546
- Lövdén, M., Fratiglioni, L., Glymour, M. M., Lindenberg, U., & Tucker-Drob, E. M. (2020). Education and Cognitive Functioning Across the Life Span. *Psychol Sci Public Interest, 21*(1), 6-41. doi:10.1177/1529100620920576
- MacLeod, C. M. (1992). The Stroop task: The "gold standard" of attentional measures. *Journal of Experimental Psychology: General, 121*(1), 12.
- Magnusdottir, B. B., Haraldsson, H. M., & Sigurdsson, E. (2021). Trail Making Test, Stroop, and Verbal Fluency: Regression-Based Norms for the Icelandic Population. *Arch Clin Neuropsychol, 36*(2), 253-266. doi:10.1093/arclin/acz049
- Málišová, E., Dančík, D., Heretik, A., Abrahámová, M., Krakovská, S., Brandoburová, P., & Hajdúk, M. (2022). Slovak version of the Trail Making Test: Normative data. *Applied Neuropsychology: Adult, 29*(6), 1476-1483. Retrieved from <https://doi.org/10.1080/23279095.2021.1890596>. doi:10.1080/23279095.2021.1890596
- Martin, T. A., Hoffman, N. M., & Donders, J. (2003). Clinical utility of the trail making test ratio score. *Appl Neuropsychol, 10*(3), 163-169. doi:10.1207/s15324826an1003_05
- Martins da Silva, A., Cavaco, S., Fernandes, J., Samões, R., Alves, C., Cardoso, M., . . . Coelho, T. (2018). Age-dependent cognitive dysfunction in untreated hereditary transthyretin amyloidosis. *J Neurol, 265*(2), 299-307. doi:10.1007/s00415-017-8668-8
- McDonald, C. R., Delis, D. C., Norman, M. A., Tecoma, E. S., & Iragui-Madoz, V. J. (2005). Is impairment in set-shifting specific to frontal-lobe dysfunction? Evidence from patients with frontal-lobe or temporal-lobe epilepsy. *Journal of the International Neuropsychological Society, 11*(4), 477-481.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods, 1*(1), 30.
- Messinis, L., Nasios, G., Mougias, A., Politis, A., Zampakis, P., Tsiamaki, E., . . . Papathanasopoulos, P. (2016). Age and education adjusted normative data and discriminative validity for Rey's Auditory Verbal Learning Test in the elderly Greek population. *J Clin Exp Neuropsychol, 38*(1), 23-39. doi:10.1080/13803395.2015.1085496
- Michaud, T. L., Su, D., Siahpush, M., & Murman, D. L. (2017). The risk of incident mild cognitive impairment and progression to dementia considering mild cognitive impairment subtypes. *Dementia and geriatric cognitive disorders extra, 7*(1), 15-29.
- Milde, S. H. (2015, June 29). *LCBC named as a world-class research group*. Retrieved April 27, 2023, from <https://www.sv.uio.no/psi/english/research/news-and-events/news/named-world-leading-research-environment.html>
- Miskin, N., Thesen, T., Barr, W. B., Butler, T., Wang, X., Dugan, P., . . . Blackmon, K. (2016). Prefrontal lobe structural integrity and trail making test, part B: converging findings from surface-based cortical thickness and voxel-based lesion symptom analyses. *Brain Imaging and Behavior, 10*(3), 675-685. Retrieved from <https://doi.org/10.1007/s11682-015-9455-8>. doi:10.1007/s11682-015-9455-8
- Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment*: Oxford University Press.
- Molinuevo, J. L., Gómez-Anson, B., Monte, G. C., Bosch, B., Sánchez-Valle, R., & Rami, L. (2011). Neuropsychological profile of prodromal Alzheimer's disease (Prd-AD) and their radiological correlates. *Archives of Gerontology and Geriatrics, 52*(2), 190-196.

- Retrieved from
<https://www.sciencedirect.com/science/article/pii/S0167494310000919>.
 doi:<https://doi.org/10.1016/j.archger.2010.03.016>
- Moll, J., de Oliveira-Souza, R., Moll, F. T., Bramati, I. E., & Andreiuolo, P. A. (2002). The cerebral correlates of set-shifting: an fMRI study of the trail making test. *Arq Neuropsiquiatr*, *60*(4), 900-905. doi:10.1590/s0004-282x2002000600002
- Montez, J. K., Hummer, R. A., & Hayward, M. D. (2012). Educational attainment and adult mortality in the United States: a systematic analysis of functional form. *Demography*, *49*(1), 315-336. doi:10.1007/s13524-011-0082-8
- Morris, G. P., Clark, I. A., & Vissel, B. (2018). Questions concerning the role of amyloid- β in the definition, aetiology and diagnosis of Alzheimer's disease. *Acta neuropathologica*, *136*, 663-689.
- Nell, V. (1999). *Cross-cultural neuropsychological assessment: Theory and practice*: Psychology Press.
- Nielsen, H., Knudsen, L., & Daugbjerg, O. (1989). Normative data for eight neuropsychological tests based on a Danish sample. *Scand J Psychol*, *30*(1), 37-45. doi:10.1111/j.1467-9450.1989.tb01066.x
- Nisbet, R., Elder, J., & Miner, G. D. (2009). *Handbook of statistical analysis and data mining applications*: Academic press.
- Nyberg, L., Magnussen, F., Lundquist, A., Baaré, W., Bartrés-Faz, D., Bertram, L., . . . Ebmeier, K. (2021). Educational attainment does not influence brain aging. *Proceedings of the National Academy of Sciences*, *118*(18), e2101644118.
- O'Connell, M. E., & Tuokko, H. (2010). Age Corrections and Dementia Classification Accuracy. *Arch Clin Neuropsychol*, *25*(2), 126-138. doi:10.1093/arclin/acp111
- Ojeda, N., Aretouli, E., Peña, J., & Schretlen, D. J. (2016). Age differences in cognitive performance: A study of cultural differences in Historical Context. *Journal of Neuropsychology*, *10*(1), 104-115. Retrieved from <https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/jnp.12059>. doi:<https://doi.org/10.1111/jnp.12059>
- Oosterhuis, H. (2017). Regression-based norming for psychological tests and questionnaires.
- Oosterhuis, H. E., van der Ark, L. A., & Sijtsma, K. (2016). Sample size requirements for traditional and regression-based norms. *Assessment*, *23*(2), 191-202.
- Oosterhuis, H. E. M., van der Ark, L. A., & Sijtsma, K. (2017). Standard Errors and Confidence Intervals of Norm Statistics for Educational and Psychological Tests. *Psychometrika*, *82*(3), 559-588. Retrieved from <https://doi.org/10.1007/s11336-016-9535-8>. doi:10.1007/s11336-016-9535-8
- Pa, J., Possin, K. L., Wilson, S. M., Quitania, L. C., Kramer, J. H., Boxer, A. L., . . . Johnson, J. K. (2010). Gray matter correlates of set-shifting among neurodegenerative disease, mild cognitive impairment, and healthy older adults. *Journal of the International Neuropsychological Society*, *16*(4), 640-650. Retrieved from <https://www.cambridge.org/core/article/gray-matter-correlates-of-setshifting-among-neurodegenerative-disease-mild-cognitive-impairment-and-healthy-older-adults/27569932F6114A06D9A6129CFDC3A73F>. doi:10.1017/S1355617710000408
- Parmenter, B. A., Testa, S. M., Schretlen, D. J., Weinstock-Guttman, B., & Benedict, R. H. (2010). The utility of regression-based norms in interpreting the minimal assessment of cognitive function in multiple sclerosis (MACFIMS). *Journal of the International Neuropsychological Society*, *16*(1), 6-16.
- Partington, J. E., & Leiter, R. G. (1949). Partington's Pathways Test. *Psychological Service Center Journal*.

- Pauker, J. D. (1988). Constructing overlapping cell tables to maximize the clinical usefulness of normative test data: Rationale and an example from neuropsychology. *Journal of clinical psychology, 44*(6), 930-933.
- Pellas, J., & Damberg, M. (2021). Assessment of executive functions in older adults: Translation and initial validation of the Swedish version of the Frontal Assessment Battery, FAB-Swe. *Appl Neuropsychol Adult, 1-5*.
doi:10.1080/23279095.2021.1990929
- Petersen, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of internal medicine, 256*(3), 183-194.
- Petersen, R. C., Roberts, R. O., Knopman, D. S., Boeve, B. F., Geda, Y. E., Ivnik, R. J., . . . Jack, C. R. (2009). Mild cognitive impairment: ten years later. *Archives of neurology, 66*(12), 1447-1455.
- Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of neurology, 56*(3), 303-308.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Van Willigen, B., & Maintainer, R. (2017). Package ‘nlme’. *Linear and nonlinear mixed effects models, version, 3*(1), 274.
- Polit, D. F. (2014). Getting serious about test–retest reliability: a critique of retest research and some recommendations. *Quality of Life Research, 23*(6), 1713-1720. Retrieved from <https://doi.org/10.1007/s11136-014-0632-9>. doi:10.1007/s11136-014-0632-9
- Poreh, A., Tolfo, S., Krivenko, A., & Teaford, M. (2017). Base-rate data and norms for the Rey Auditory Verbal Learning Embedded Performance Validity Indicator. *Applied Neuropsychology: Adult, 24*(6), 540-547. Retrieved from <https://doi.org/10.1080/23279095.2016.1223670>.
doi:10.1080/23279095.2016.1223670
- Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*(1), 33-65.
- Raudeberg, R., L. Iverson, G., & Hammar, Å. (2019). Norms matter: U.S. normative data under-estimate cognitive deficits in Norwegians with schizophrenia spectrum disorders. *The Clinical Neuropsychologist, 33*(sup1), 58-74. Retrieved from <https://doi.org/10.1080/13854046.2019.1590641>.
doi:10.1080/13854046.2019.1590641
- Regard, M. (1983). COGNITIVE RIGIDITY AND FLEXIBILITY: A NEUROPSYCHOLOGICAL STUDY.
- Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation* (Vol. 4): Reitan Neuropsychology.
- Reitan, R. M., & Wolfson, D. (1994). A selective and critical review of neuropsychological deficits and the frontal lobes. *Neuropsychology review, 4*, 161-198.
- Revelle, W., & Revelle, M. W. (2015). Package ‘psych’. *The comprehensive R archive network, 337*, 338.
- Rey, A. (1958). L'examen clinique en psychologie.
- Ricci, M., Ruggeri, M., Gnisci, C., Pizzoni, L., Gerace, C., & Blundo, C. (2022). Improving Amnesia Diagnostic Accuracy with RAVLT Single Scores and Composite Indices: Italian Normative Data. *Arch Clin Neuropsychol, 37*(8), 1749-1764.
doi:10.1093/arclin/acac055
- Ritchie, S. J., & Tucker-Drob, E. M. (2018). How Much Does Education Improve Intelligence? A Meta-Analysis. *Psychol Sci, 29*(8), 1358-1369.
doi:10.1177/0956797618774253

- Rivera, D., Perrin, P. B., Stevens, L. F., Garza, M. T., Weil, C., Saracho, C. P., . . . Arango-Lasprilla, J. C. (2015). Stroop Color-Word Interference Test: Normative data for the Latin American Spanish speaking adult population. *NeuroRehabilitation*, *37*, 591-624. doi:10.3233/NRE-151281
- Rodríguez-Lorenzana, A., Ramos-Usuga, D., Díaz, L. A., Mascialino, G., Yacelga Ponce, T., Rivera, D., & Arango-Lasprilla, J. C. (2021). Normative data of neuropsychological tests of attention and executive functions in Ecuadorian adult population. *Aging, Neuropsychology, and Cognition*, *28*(4), 508-527. Retrieved from <https://doi.org/10.1080/13825585.2020.1790493>. doi:10.1080/13825585.2020.1790493
- Rohling, M. L., Miller, R. M., Axelrod, B. N., Wall, J. R., Lee, A. J., & Kinikini, D. T. (2015). Is co-norming required? *Archives of Clinical Neuropsychology*, *30*(7), 611-633.
- Ryder, T. (2021). Testkvalitetsprosjektet-del 1: Norske psykologers testholdninger og testbruk. *Tidsskrift for Norsk psykologforening*, *58*(1), 28-37.
- Salthouse, T. A. (1993). Speed mediation of adult age differences in cognition. *Developmental Psychology*, *29*(4), 722-738. doi:10.1037/0012-1649.29.4.722
- Salthouse, T. A. (2010). Selective review of cognitive aging. *J Int Neuropsychol Soc*, *16*(5), 754-760. doi:10.1017/S1355617710000706
- Samordna opptak. (2023). Studieoversikten. Retrieved from <https://sok.samordnaopptak.no/#/admission/12/studies>
- Sánchez-Cubillo, I., Periáñez, J. A., Adrover-Roig, D., Rodríguez-Sánchez, J. M., Ríos-Lago, M., Tirapu, J., & Barceló, F. (2009). Construct validity of the Trail Making Test: role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *Journal of the International Neuropsychological Society*, *15*(3), 438-450.
- Saury, J. M., & Emanuelson, I. (2017). Neuropsychological Assessment of Hippocampal Integrity. *Appl Neuropsychol Adult*, *24*(2), 140-151. doi:10.1080/23279095.2015.1113536
- Savla, G. N., Twamley, E. W., Thompson, W. K., Delis, D. C., Jeste, D. V., & Palmer, B. W. (2011). Evaluation of specific executive functioning skills and the processes underlying executive control in schizophrenia. *Journal of the International Neuropsychological Society*, *17*(1), 14-23.
- Schmidt, M. (1996). *Rey auditory verbal learning test: A handbook* (Vol. 17): Western Psychological Services Los Angeles, CA.
- Schneider, E. C., Shah, A., Doty, M. M., Tikkanen, R., Fields, K., Williams, R., & II, M. M. (2021). Reflecting Poorly: Health Care in the US Compared to Other High-Income Countries. *New York: The Commonwealth Fund*.
- Selander, H., Wressle, E., & Samuelsson, K. (2020). Cognitive prerequisites for fitness to drive: Norm values for the TMT, UFOV and NorSDSA tests. *Scand J Occup Ther*, *27*(3), 231-239. doi:10.1080/11038128.2019.1614214
- Senior, G., Piovesana, A., & Beaumont, P. (2018). Discrepancy analysis and Australian norms for the Trail Making Test. *The Clinical Neuropsychologist*, *32*(3), 510-523.
- Sherman, E. M., Brooks, B. L., Iverson, G. L., Slick, D. J., & Strauss, E. (2010). Reliability and validity in neuropsychology. In *The little black book of neuropsychology: A syndrome-based approach* (pp. 873-892): Springer.
- Skirbekk, V., Loichinger, E., & Weber, D. (2012). Variation in cognitive functioning as a refined approach to comparing aging across countries. *Proc Natl Acad Sci U S A*, *109*(3), 770-774. doi:10.1073/pnas.1112173109
- Skirbekk, V., Stonawski, M., Bonsang, E., & Staudinger, U. M. (2013). The Flynn effect and population aging. *Intelligence*, *41*(3), 169-177.

- Slaughter, S., Cole, D., Jennings, E., & Reimer, M. A. (2007). Consent and assent to participate in research from people with dementia. *Nurs Ethics*, *14*(1), 27-40. doi:10.1177/0969733007071355
- Sliwinski, M., Lipton, R. B., Buschke, H., & Stewart, W. (1996). The Effects of Preclinical Dementia on Estimates of Normal Cognitive Functioning in Aging. *The Journals of Gerontology: Series B*, *51B*(4), P217-P225. Retrieved from <https://doi.org/10.1093/geronb/51B.4.P217>. doi:10.1093/geronb/51B.4.P217
- Sparding, T., Silander, K., Pålsson, E., Östlind, J., Sellgren, C., Ekman, C. J., . . . Landén, M. (2015). Cognitive functioning in clinically stable patients with bipolar disorder I and II. *PLoS One*, *10*(1), e0115562. doi:10.1371/journal.pone.0115562
- Specka, M., Weimar, C., Stang, A., Jöckel, K. H., Scherbaum, N., Hoffmann, S. S., . . . Jokisch, M. (2022). Trail Making Test Normative Data for the German Older Population. *Arch Clin Neuropsychol*, *37*(1), 186-198. doi:10.1093/arclin/acab027
- St-Hilaire, A., Parent, C., Potvin, O., Bherer, L., Gagnon, J. F., Joubert, S., . . . Macoir, J. (2018). Trail Making Tests A and B: regression-based normative data for Quebec French-speaking mid and older aged adults. *Clin Neuropsychol*, *32*(sup1), 77-90. doi:10.1080/13854046.2018.1470675
- Stallings, G., Boake, C., & Sherer, M. (1995). Comparison of the California Verbal Learning Test and the Rey Auditory Verbal Learning Test in head-injured patients. *J Clin Exp Neuropsychol*, *17*(5), 706-712. doi:10.1080/01688639508405160
- Statistics Norway. (2023a). Innvandrere og norskfødte med innvandrerforeldre. Retrieved from <https://www.ssb.no/befolkning/innvandrere/statistikk/innvandrere-og-norskfodte-med-innvandrereforeldre>
- Statistics Norway. (2023b). Lang utdanning gir ikke nødvendigvis høyere inntekt gjennom livet. Retrieved from <https://www.ssb.no/inntekt-og-forbruk/inntekt-og-formue/artikler/lang-utdanning-gir-ikke-nodvendigvis-hoyere-inntekt-gjennom-livet>
- Steinberg, B. A., Bieliauskas, L. A., Smith, G. E., & Ivnik, R. J. (2005). Mayo's older Americans normative studies: age-and IQ-adjusted norms for the trail-making test, the stroop test, and MAE controlled oral word association test. *The Clinical Neuropsychologist*, *19*(3-4), 329-377.
- Stern, Y., Albert, M., Barnes, C. A., Cabeza, R., Pascual-Leone, A., & Rapp, P. R. (2023). A framework for concepts of reserve and resilience in aging. *Neurobiology of Aging*, *124*, 100-103.
- Storsve, A. B., Fjell, A. M., Tamnes, C. K., Westlye, L. T., Overbye, K., Aasland, H. W., & Walhovd, K. B. (2014). Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: regions of accelerating and decelerating change. *J Neurosci*, *34*(25), 8488-8498. doi:10.1523/jneurosci.0391-14.2014
- Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*: American chemical society.
- Stricker, N. H., Christianson, T. J., Lundt, E. S., Alden, E. C., Machulda, M. M., Fields, J. A., . . . Mielke, M. M. (2021). Mayo normative studies: regression-based normative data for the auditory verbal learning test for ages 30–91 years and the importance of adjusting for sex. *Journal of the International Neuropsychological Society*, *27*(3), 211-226.
- Strobel, C., Johansen, H., Aga, O., Bekkhus-Wetterberg, P., Brierly, M., Egeland, J., . . . Schanke, A. (2018). Manual Norsk Revidert trail making test (TMT-NR3). In: Retrieved 12/08/2021 from <https://www.aldringoghelse.no/ah-archive> . . .
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.

- Stuss, D. T., Floden, D., Alexander, M., Levine, B., & Katz, D. (2001). Stroop performance in focal lesion patients: dissociation of processes and frontal lobe lesion location. *Neuropsychologia*, *39*(8), 771-786.
- Sundet, J. M., Barlaug, D. G., & Tojussen, T. M. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence (Norwood)*, *32*(4), 349-362. doi:10.1016/s0160-2896(04)00052-2
- Suzuki, H., Sakuma, N., Kobayashi, M., Ogawa, S., Inagaki, H., Eda, H., . . . Awata, S. (2022). Normative Data of the Trail Making Test Among Urban Community-Dwelling Older Adults in Japan. *Front Aging Neurosci*, *14*, 832158. doi:10.3389/fnagi.2022.832158
- Tamnes, C. K., Walhovd, K. B., Dale, A. M., Østby, Y., Grydeland, H., Richardson, G., . . . Fjell, A. M. (2013). Brain development and aging: overlapping and unique patterns of change. *Neuroimage*, *68*, 63-74. doi:10.1016/j.neuroimage.2012.11.039
- Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn Effect in reverse. *Personality and Individual Differences*, *39*(4), 837-843.
- Torres, I. J., Flashman, L. A., O'leary, D. S., & Andreasen, N. C. (2001). Effects of retroactive and proactive interference on word list recall in schizophrenia. *Journal of the International Neuropsychological Society*, *7*(4), 481-490.
- Uttl, B. (2005). Measurement of individual differences: lessons from memory assessment in research and clinical practice. *Psychol Sci*, *16*(6), 460-467. doi:10.1111/j.0956-7976.2005.01557.x
- Vakil, E., Greenstein, Y., & Blachstein, H. (2010). Normative Data for Composite Scores for Children and Adults Derived from the Rey Auditory Verbal Learning Test. *The Clinical Neuropsychologist*, *24*(4), 662-677. Retrieved from <https://doi.org/10.1080/13854040903493522>. doi:10.1080/13854040903493522
- Van Breukelen, G. J., & Vlaeyen, J. W. (2005). Norming clinical questionnaires with multiple regression: the Pain Cognition List. *Psychol Assess*, *17*(3), 336-344. doi:10.1037/1040-3590.17.3.336
- van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *European Review of Applied Psychology*, *54*(2), 119-135. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1162908803000057>. doi:<https://doi.org/10.1016/j.erap.2003.12.004>
- Van der Elst, W., Molenberghs, G., van Tetering, M., & Jolles, J. (2017). Establishing normative data for multi-trial memory tests: the multivariate regression-based approach. *Clin Neuropsychol*, *31*(6-7), 1173-1187. doi:10.1080/13854046.2017.1294202
- Van der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., & Jolles, J. (2006). 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*(1), 62-79.
- Varjadic, A., Mantini, D., Demeyere, N., & Gillebert, C. R. (2018). Neural signatures of Trail Making Test performance: Evidence from lesion-mapping and neuroimaging studies. *Neuropsychologia*, *115*, 78-87. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0028393218301246>. doi:<https://doi.org/10.1016/j.neuropsychologia.2018.03.031>
- Vaskinn, A., & Egeland, J. (2012). Testbruksundersøkelsen: En oversikt over tester brukt av norske psykologer. *Tidsskrift for Norsk psykologforening*, *49*(7), 658-665.
- Verbeke, G., & Molenberghs, G. (2009). *Linear Mixed Models for Longitudinal Data*: Springer Science & Business Media.

- Vicente, S. G., Rivera, D., Barbosa, F., Gaspar, N., Dores, A. R., Mascialino, G., & Arango-Lasprilla, J. C. (2021). Normative data for tests of attention and executive functions in a sample of European Portuguese adult population. *Aging, Neuropsychology, and Cognition*, 28(3), 418-437. Retrieved from <https://doi.org/10.1080/13825585.2020.1781768>. doi:10.1080/13825585.2020.1781768
- Vogel, A., Stokholm, J., & Jørgensen, K. (2012). Performances on Rey Auditory Verbal Learning Test and Rey Complex Figure Test in a healthy, elderly Danish sample--reference data and validity issues. *Scand J Psychol*, 53(1), 26-31. doi:10.1111/j.1467-9450.2011.00909.x
- Vogel, A., Stokholm, J., & Jørgensen, K. (2013). Performances on Symbol Digit Modalities Test, Color Trails Test, and modified Stroop test in a healthy, elderly Danish sample. *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn*, 20(3), 370-382. doi:10.1080/13825585.2012.725126
- Voncken, L., Albers, C. J., & Timmerman, M. E. (2019). Improving confidence intervals for normed test scores: Include uncertainty due to sampling variability. *Behavior Research Methods*, 51(2), 826-839. Retrieved from <https://doi.org/10.3758/s13428-018-1122-8>. doi:10.3758/s13428-018-1122-8
- Voncken, L., Albers, C. J., & Timmerman, M. E. (2019). Model selection in continuous test norming with GAMLSS. *Assessment*, 26(7), 1329-1346.
- Vuoksimaa, E., McEvoy, L. K., Holland, D., Franz, C. E., & Kremen, W. S. (2020). Modifying the minimum criteria for diagnosing amnesic MCI to improve prediction of brain atrophy and progression to Alzheimer's disease. *Brain Imaging Behav*, 14(3), 787-796. doi:10.1007/s11682-018-0019-6
- Wagle, J., Selbæk, G., Benth, J. Š., GjØra, L., Rønqvist, T. K., Bekkhus-Wetterberg, P., . . . Engedal, K. (2023). The CERAD Word List Memory Test: Normative Data Based on a Norwegian Population-Based Sample of Healthy Older Adults 70 Years and Above. The HUNT Study. *J Alzheimers Dis*, 91(1), 321-343. doi:10.3233/JAD-220672
- Wallin, A., Nordlund, A., Jonsson, M., Lind, K., Edman, Å., Göthlin, M., . . . Börjesson-Hanson, A. (2016). The Gothenburg MCI study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow & Metabolism*, 36(1), 114-131.
- Warrington, E. K. (1991). Visual object and space perception battery. (*No Title*).
- Weber, D., Skirbekk, V., Freund, I., & Herlitz, A. (2014). Changing face of cognitive gender differences in Europe. *Proc Natl Acad Sci U S A*, 111(32), 11673-11678. doi:10.1073/pnas.1319538111
- Wei, M., Shi, J., Li, T., Ni, J., Zhang, X., Li, Y., . . . Tian, J. (2018). Diagnostic Accuracy of the Chinese Version of the Trail-Making Test for Screening Cognitive Impairment. *J Am Geriatr Soc*, 66(1), 92-99. doi:10.1111/jgs.15135
- Willse, J. T. (2022). CTT: Classical Test Theory Functions. R package version 2.3.3. URL: <https://rdrr.io/cran/CTT/>
- Williams, R. L. (2013). Overview of the Flynn effect. *Intelligence*, 41(6), 753-764.
- Wimo, A., Seeher, K., Cataldi, R., Cyhlarova, E., Dieleman, J. L., Frisell, O., . . . Dua, T. (2023). The worldwide costs of dementia in 2019. *Alzheimers Dement*. doi:10.1002/alz.12901
- Woodard, J. L. (2006). Memory performance indexes for the Rey Auditory Verbal Learning Test. In *The quantified process approach to neuropsychological assessment*. (pp. 105-141). Philadelphia, PA, US: Taylor & Francis.

- Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Leirer, V. O. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res*, *17*(1), 37-49.
- Yesavage, J. A., & Sheikh, J. I. (1986). 9/Geriatric depression scale (GDS) recent evidence and development of a shorter version. *Clinical gerontologist*, *5*(1-2), 165-173.
- Yu, L., Boyle, P. A., Segawa, E., Leurgans, S., Schneider, J. A., Wilson, R. S., & Bennett, D. A. (2015). Residual decline in cognition after adjustment for common neuropathologic conditions. *Neuropsychology*, *29*(3), 335.
- Zachary, R. A., & Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of clinical psychology*, *41*(1), 86-94.
- Aarts, E., & Oosterhuis, H. E. M. Regression-Based Norming for Psychological Tests and Questionnaires.

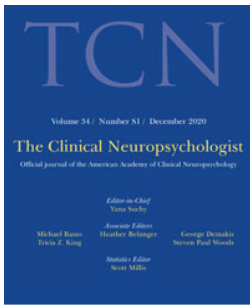
9. Papers I-III

Paper 1

Espenes, J., Hessen, E., Eliassen, I.V., Waterloo, K., Eckerström, M., Sando, S.B., ... & Kirsebom, B.E. (2020).

Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years.

The Clinical Neuropsychologist, 34(sup1), 110-126.



Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years

Jacob Espenes, Erik Hessen, Ingvild Vøllo Eliassen, Knut Waterloo, Marie Eckerström, Sigrid Botne Sando, Santiago Timón, Anders Wallin, Tormod Fladby & Bjørn-Eivind Kirsebom

To cite this article: Jacob Espenes, Erik Hessen, Ingvild Vøllo Eliassen, Knut Waterloo, Marie Eckerström, Sigrid Botne Sando, Santiago Timón, Anders Wallin, Tormod Fladby & Bjørn-Eivind Kirsebom (2020) Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years, *The Clinical Neuropsychologist*, 34:sup1, 110-126, DOI: [10.1080/13854046.2020.1829068](https://doi.org/10.1080/13854046.2020.1829068)

To link to this article: <https://doi.org/10.1080/13854046.2020.1829068>



© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 09 Oct 2020.



[Submit your article to this journal](#)



Article views: 3351



[View related articles](#)













[View Crossmark data](#)



Citing articles: 8 [View citing articles](#)

Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years

Jacob Espenes^{a,b} , Erik Hossen^{c,d} , Ingvild Vølle Eliassen^{c,d} , Knut Waterloo^{a,b} , Marie Eckerström^e , Sigrid Botne Sando^{f,g} , Santiago Timón^{c,h} , Anders Wallin^e , Tormod Fladby^{c,i}  and Bjørn-Eivind Kirsebom^{a,b} 

^aDepartment of Psychology, Faculty of Health Sciences, The Arctic University of Norway, Tromsø, Norway; ^bDepartment of Neurology, University Hospital of North Norway, Tromsø, Norway; ^cDepartment of Neurology, Akershus University Hospital, Lørenskog, Norway; ^dDepartment of Psychology, University of Oslo, Oslo, Norway; ^eDepartment of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; ^fDepartment of Neuromedicine and Movement Science, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway; ^gDepartment of Neurology and Clinical Neurophysiology, University Hospital of Trondheim, Trondheim, Norway; ^hDepartamento de Inteligencia Artificial, Universidad Nacional de Educación a Distancia, Madrid, Spain; ⁱInstitute of Clinical Medicine, Campus Ahus, University of Oslo, Oslo, Norway

ABSTRACT

Objective: The trail making test (TMT) is one of the most widely used neuropsychological tests. TMT-A provides measures of visual scanning/visuomotor speed and TMT-B involves additional demands on executive functions. Derived scores TMT B-A and TMT B/A enhance measures of executive functioning. However, simple B-A subtraction may lead to false estimates of executive dysfunction in clinical samples. Norms for TMT have been published in several countries but are currently lacking for Scandinavia.

Methods: A total of 292 healthy controls between age 41 and 84 years were included from the Norwegian “Dementia Disease Initiation” (DDI) study ($n=170$) and the Gothenburg Mild Cognitive Impairment (MCI) study ($n=122$). We used a regression-based procedure to develop demographically adjusted norms for basic (TMT-A and TMT-B) and derived measures (TMT B-A and B/A). We also propose a regression-based alternative to the TMT B-A measure named “TMT- β ”. The proposed norms were compared to norms from Heaton et al. and Tombaugh.

Results: Due to differences in the estimated normative effects of demographics on performance, the proposed norms for TMT were better suited in the Scandinavian sample compared with published non-Scandinavian norms. The proposed TMT- β measure was highly correlated to TMT B-A ($r=0.969$, $p<0.001$).

Conclusion: We here propose demographically adjusted norms for the TMT for ages 41 through 84 years based on a Scandinavian

ARTICLE HISTORY

Received 15 May 2020
Accepted 19 September 2020
Published online 9 October 2020

KEYWORDS

Neuropsychological tests; norms; trail making test; TMT; Scandinavia; Norway; Sweden; cross-cultural neuropsychology

CONTACT Jacob Espenes  johan.j.espenes@uit.no  Department of Psychology, Faculty of Health Sciences, The Arctic University of Norway, Hansine Hansens veg 18, Tromsø 9019, Norway.

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

sample. We also present the regression-based derived measure TMT- β which may resolve issues with the conventional TMT B-A measure.

Introduction

Performance on the trail making test (TMT) is mediated through a set of global neural mechanisms (Moll et al., 2002) and TMT is sensitive to a variety of conditions with neurological deficits (Gonçalves et al., 2013). TMT is therefore suitable as a screening tool for neurological integrity and identification of individuals in need of cognitive assessment (Reitan & Wolfson, 2004). Basic task demands of TMT-A and TMT-B are visual search and/or visuomotor speed. TMT-B is a more difficult task, involving additional demands on executive functions including working memory and cognitive flexibility due to the alternation between numbers and letters (Sanchez-Cubillo et al., 2009). Derived measures of TMT have been suggested to highlight measurements of executive functions associated with TMT-B, primarily difference score TMT B-A (Lezak et al., 2012, p. 423) and ratio score TMT B/A (Arbuthnott & Frank, 2000; Lamberty et al., 1994).

Clinicians rely on published norms which aim to correct for demographics known to influence test performance. On TMT, increasing age is associated with decreased performance (Gonçalves et al., 2013; Goul & Brown, 1970; Kennedy, 1981; Stuss et al., 1988) and higher educational attainment relates to increases in performance especially on TMT-B (Heaton et al., 2004; Peña-Casanova et al., 2009; Periañez et al., 2007; Tombaugh, 2004). Derived measures TMT B-A and TMT B/A are less affected by variations in age and education compared with the basic measures (Gonçalves et al., 2013). Most studies do not find sex differences on TMT (Mitrushina et al., 2005, p. 69). Normative studies investigating the effects of age and educational attainment on TMT scores show varying results due to differences in sample characteristics (e.g. range of educational attainment and age in the sample) and may limit the applicability of norms across different populations. Indeed, TMT norms have been shown to produce markedly diverging estimates when applied to different populations, ranging from 0.8 to 1.4 standard deviations (Fernandez & Marcopulos, 2008). In addition, cohort effects have been found on TMT, likely due to advancements in educational quality and health (i.e. a Flynn Effect; Dickinson & Hiscock, 2011; Dodge et al., 2014). To resolve these issues, local norms have been developed for the TMT in several countries (Abi Chahine et al., 2019; Cavaco et al., 2013; Gonçalves et al., 2013; Siciliano et al., 2019; St-Hilaire et al., 2018).

To our knowledge, test norms for TMT based on a Scandinavian sample have not been published. Thus, the first objective of this study was to investigate the influence of age, education, and sex on TMT scores in a sample of healthy Norwegians and Swedes between 41 and 84 years of age ($n=292$) and develop norms for the basic and derived measures of TMT using a regression-based norming procedure. Second, we compare the current proposed norms with two sets of norms (Heaton et al. 2004; Tombaugh, 2004) frequently applied by clinicians and researchers and recommended by Norwegian health authorities in clinical use (Strobel et al., 2018). Third, we propose an alternate method for computing the conventional TMT B-A measure which might have applications in clinical populations.

A disadvantage of the simple subtraction method TMT B-A is that an elevated difference score is interpreted as difficulties with the additional task demands of TMT-B, indicating deficits in executive functions (cognitive flexibility and working memory). However, TMT-B is also more demanding than TMT-A on visual search and/or visuo-motor abilities due to increased amount of connections to be drawn, and the distance between connections (Gaudino et al., 1995). Patients with general visuomotor and visual scanning deficits resulting in reduced performance on both TMT-A and TMT-B may therefore show a disproportionate increase in time to completion on TMT-B (Senior et al., 2018). Thus, a high TMT B-A difference could also be due to general visuomotor or visual scanning deficits rather than executive deficits. As shown by Senior et al. (2018), normative values on TMT B-A are based on mean values from the entire sample and do not accommodate this non-linear relationship by accounting for individual variability on TMT-A. We therefore propose an alternative method for the derived measure TMT B-A by regressing age and education along with scores from TMT-A on scores from TMT-B using multiple regression analysis. This approach resolves the issues with conventional B-A subtraction while simultaneously controlling for pertinent demographics. We have named this new measure “TMT- β ” to avoid confusion with the conventional TMT B-A approach.

Methods and materials

Participants

We included healthy controls from the Norwegian Dementia Disease Initiation Study (DDI; $n = 170$) and the Swedish Gothenburg mild cognitive impairment (MCI) study ($n = 122$). DDI is a national multicenter longitudinal study aimed at early detection and diagnosis of common neurodegenerative diseases such as Alzheimer’s disease (AD). Participants from DDI were recruited between January 2013 and October 2018. The Gothenburg MCI study started in 1999 and is an ongoing single-center study on early phases of AD and vascular dementia based in Sahlgrenska University Hospital in Sweden. Participants were recruited between January 2001 and March 2014.

Criteria for inclusion of healthy controls from the DDI study were ages 40 through 80, absence of subjective symptoms of cognitive decline and MMSE score >26 and a native language of Norwegian, Danish, or Swedish. Participants in the DDI cohort were recruited from all Norwegian health regions. Healthy controls were primarily recruited from spouses of symptom group participants and secondarily by self-referral through advertisements in local media and from orthopedic wards. All participants from the DDI study followed a standardized procedure for assessment following a Case Report Form (CRF) developed for DDI and is described in detail in Fladby et al. (2017). Briefly, this included standardized neurological and physical examinations by neurologist, brief neuropsychological assessment, and standardized interview involved taking a medical history from participants and informants. Licensed psychologists, neurologists, licensed study nurses, or psychologists-in-training under supervision from licensed psychologists performed cognitive assessments. Patients with history of stroke, severe psychiatric disorder including major depression, intellectual disability or developmental disorders, and severe somatic disorders that may influence cognitive functions were excluded.

Table 1. Demographics, raw scores, and *T*-scores of the healthy controls from the dementia disease initiation (DDI) and Gothenburg mild cognitive impairment (MCI) study ($n = 292$).

Variables	Test scores/demographics		t/χ^2	p
	DDI controls $n = 170$	Gothenburg MCI $n = 122$		
Age M (SD) [range]	62.0 (9.4) [41 – 84]	64.3 (6.5) [49 – 77]	$t = -2.39$	<0.05
Female n (%)	100 (58.8%)	74 (60.7%)	$\chi^2 0.10$	ns
Years of education M (SD) [range]	13.8 (3.3) [7 – 23]	12.4 (3.2) [6 – 24]	$t = 3.83$	$<.001$
TMT-A s M (SD)	35.0 (11.6)	34.8 (10.4)	$t = 0.19$	ns
TMT-B s M (SD)	82.6 (28.4)	82.2 (23.4)	$t = 0.13$	ns
TMT B-A raw score M (SD)	47.6 (25.6)	47.4 (18.5)	$t = 0.06$	ns
TMT B/A raw score M (SD)	2.5 (0.9)	2.4 (0.6)	$t = 0.53$	ns
TMT-A T -scores M (SD)	49.61 (10.2)	50.4 (9.7)	$t = -0.69$	ns
TMT-B T -scores M (SD)	49.2 (10.1)	51.0 (9.7)	$t = -0.16$	ns
TMT B-A T -scores M (SD)	49.7 (10.7)	50.3 (8.9)	$t = -0.05$	ns
TMT B/A T -scores M (SD)	49.8 (11.1)	50.3 (8.2)	$t = -0.41$	ns
TMT- β T -scores M (SD)	49.3 (10.8)	50.9 (8.7)	$t = -0.13$	ns

n , number of participants; p , p -value; t , t statistic; ns , non-significant result; χ^2 , Pearson Chi-Square. Results are presented as mean (standard deviation) [range] except for sex which is characterized by female percentage; T -scores adjusted for pertinent demographics applying current proposed norms (Table 3).

Healthy controls from the Gothenburg MCI study were primarily recruited through senior citizen organizations and a small proportion were relatives of symptom group participants. Inclusion criteria for healthy controls in the Gothenburg MCI study were age between 50 and 79, absence of subjective symptoms of cognitive decline and MMSE score >26 . Exclusion criteria were severe somatic diseases and severe psychiatric disorders, which could potentially influence cognitive performance. Neuropsychological examinations including TMT-A and TMT-B were performed by licensed clinical psychologists or psychologist-in-training under supervision by a licensed clinical psychologist. For further description of the Gothenburg MCI study cohort, see Wallin et al. (2016).

Between cohort comparisons of demographics and cognitive performance

Demographics and raw scores on basic and derived measures for DDI ($n = 170$) and Gothenburg MCI study ($n = 122$) controls are compared in Table 1. Although participants from the Gothenburg MCI study were older ($p < 0.05$) and had less education ($p < 0.001$) compared to the DDI controls, no differences were observed between cohorts for basic or derived TMT raw scores or T -scores adjusted for pertinent demographics. Due to large differences in time of inclusion within the Gothenburg MCI cohort (i.e. participants included within a 13-year time frame), potential cohort effects were investigated by including a separate variable accounting for time of testing on TMT-A and TMT-B T -scores. Results from this analysis showed that time of testing was not a significant predictor of performance on TMT.

TMT administration

The TMT (Reitan & Wolfson, 1985) was administered following standardized instructions described in Strauss et al. (2006, pp. 656–657). Reitan and Wolfson (1985) version

Table 2. Raw score to scaled score conversions.

Scaled score	TMT-A	TMT-B	TMT B – TMT A	TMT B/TMT A	Scaled score
1					1
2	≥71	≥166	≥131	≥5.330	2
3	66–70	160–165	121–130	4.810–5.329	3
4	64–65	155–159	106–120	4.380–4.809	4
5	58–63	135–154	99–105	3.980–4.379	5
6	54–57	121–134	80–98	3.600–3.979	6
7	46–53	108–120	68–79	3.180–3.599	7
8	40–45	96–107	59–67	2.830–3.179	8
9	37–39	86–95	50–58	2.580–2.829	9
10	33–36	78–85	43–49	2.350–2.579	10
11	29–32	71–77	38–42	2.100–2.349	11
12	27–28	63–70	32–37	1.930–2.099	12
13	25–26	58–62	28–31	1.810–1.929	13
14	22–24	51–57	23–27	1.630–1.809	14
15	21	47–50	20–22	1.540–1.629	15
16	19–20	41–46	10–19	1.330–1.539	16
17	17–18	40–41	2–10	1.040–1.329	17
18	16	34–39	(–4)–1	0.890–1.039	18
19	≤15	≤33	≤(–5)	≤0.889	19

Conversions were performed to normalize TMT scores from healthy controls ($n = 292$). Normalized scaled scores were later used for development of normative models (Table 3).

of TMT is administered in two parts: In TMT-A, the participant is required to connect 25 encircled numbers from low to high, while in TMT-B, the participant must alternate between numbers and letters, from low to high (e.g. 1-A-2-B-3-C). Scoring criteria is time to completion, measured manually by digital stopwatch. In short, participants were asked to complete the task as quickly as they could without making mistakes and were presented with a rehearsal trial before the test. Participants were given a moment to familiarize with initial connections and finishing point. Time to completion was recorded between the initiation of the first pen stroke and terminated at completion of the task. In case of mistakes (e.g. connecting wrong number to letter), the participants were corrected by the administrator and promptly guided to the last correctly connected letter or number. Time was not paused during this correction. If a participant aborted TMT-B, maximum time to completion was set (300 s), although no participants in the healthy control groups achieved maximum time nor were reported to abort the assignment. In the normative sample $n = 1$ participant (0.34%) only had available data from TMT-A and was excluded from analysis.

Data analysis

Regression norming procedure

Following procedures outlined in Kirsebom et al. (2019) and Testa et al. (2009) regression-based norms were developed based on the normative performance of the included healthy controls ($n = 292$). To normalize measures of the TMT, we first determined the reverse cumulative frequency distribution for TMT raw scores (i.e. the scaled score distributions were reversed to ensure that higher times to completion was equal to lower performance in our normative models), and then converted raw scores into standardized scaled scores ($M = 10$, $SD = 3$). Multiple linear regression analyses were conducted on the standardized scaled scores (Table 2) from basic and

Table 3. Normative regression models for the TMT in healthy controls ($n = 292$).

Variable	Predictor	b	Standard error b	t	p	Partial r^2	Adjusted r^2	SD residual
TMT-A	Intercept	19.437	1.188	16.36	<0.001			2.675
	Age	-0.144	0.019	-7.70	<0.001	0.170	0.167	
TMT-B	Intercept	16.921	1.430	11.84	<0.001			2.645
	Age	-0.139	0.019	-7.40	<0.001	0.159		
TMT-B (age only)	Intercept	19.854	1.201	16.53	<0.001			2.704
	Age	-0.150	0.019	-7.93	<0.001	0.178	0.175	
TMT B-A	Intercept	13.844	1.496	9.26	<0.001			2.767
	Age	-0.091	0.020	-4.61	<0.001	0.069		
TMT B-A (age only)	Intercept	16.855	1.256	13.42	<0.001			2.827
	Age	-0.102	0.020	-5.15	<0.001	0.084	0.081	
TMT B/A	Intercept	8.705	0.705	12.34	<0.001			2.958
	Education	0.141	0.052	2.73	<0.01	0.025	0.022	
TMT- β	Intercept	8.475	1.600	5.30	<0.001			2.352
	Age	-0.075	0.018	-4.12	<0.001	0.055		
	Education	0.149	0.042	3.56	<0.001	0.042		
TMT- β (age only)	TMT-A	0.453	0.052	8.73	<0.001	0.209	0.372	2.403
	Intercept	10.851	1.483	7.32	<0.001			
	Age	-0.083	0.018	-4.50	<0.001	0.066		
	TMT-A	0.463	0.053	8.76	<0.001	0.210	0.346	

Regression analyses were performed on normalized scaled scores (Table 2). b , unstandardized regression coefficient; t , the t -test statistic; SD Residual, standard deviation of the residual; p , p -value; partial r^2 , explained variance from individual predictor; adjusted r^2 , combined explained variance from the model; standard error b , standard error of the unstandardized beta coefficient.

derived measures of the TMT in a healthy control group ($n = 292$) with age, sex, and education included as predictors. We included squared and interaction terms in our models to investigate potential non-linear effects of age (i.e. performance on TMT increasing at younger ages, then dropping off at older ages), and potential interaction effects between predictors such as between age and education, sex and education, as well as three-way interaction effects between age, sex, and education. For the proposed TMT- β measure, we included the normalized scaled scores for the TMT-A as a covariate.

All measures of basic and derived TMT scores were analyzed using a backwards regression method and only models with predictors that significantly contributed to the overall explained variance were selected. We found that the Gothenburg MCI study cohort was older and less educated (potentially due to differences in recruitment methods) and we therefore included a covariate to assess a potential difference between cohorts for the TMT measures. However, when controlling for demographics, no differences between cohorts were observed. There were no effects of sex on performance for any of the measures. For TMT-A, only age remained a significant predictor of test performance. For the TMT-B and derived measures TMT B-A and TMT- β , both age and education were significant predictors. For TMT B/A, only education significantly predicted performance. On the proposed measure TMT- β , age, education, and normalized scaled scores on TMT-A were significant predictors. None of the squared terms or interaction terms provided additional explained variance in the model. Education may not always be a relevant normative demographic for all target populations (e.g. low educational attainment while scoring above average on age adjusted measures of intelligence). Thus, we also provide regression-norms omitting education as a covariate. These norms may be applied to scores from individuals who

did not have access to education but otherwise would have benefited from it, or when deemed appropriate by the clinician. Regression coefficients and partial r^2 values for the different predictors are presented in Table 3. For these models, we assessed plots of regression predicted values to residuals values to ensure that the assumption of homoscedasticity was not violated, and normality of the residuals were visually inspected with Q-Q plots. No collinearity between predictor variables were observed in the selected models (variance inflation factor <1.2).

Calculating normative performance using regression-based norms

The normative effects of demographics on performance are first determined using the regression coefficients obtained from the multiple regression analysis (Table 3) described above using the following formula (Intercept + [individual age*age coefficient] + [years of education * education coefficient]). For example, for a 60-year-old woman with 13 years of education, the resulting equation on TMT-B would be: $[(16.921) + [60 * -0.139]] + (13 * 0.170)$. This formula produces an individual predicted scaled score for TMT-B. We then subtract the scaled score obtained by the individual (Table 2) from the demographically adjusted predicted scaled score and divide by the standard deviation of the regression model residuals (Table 3) which yields a standardized Z-score (Obtained scaled model score – predicted scaled score/standard deviation of the residuals obtained from the regression = Z-score). The resulting Z-score is the demographically adjusted normative score based on the healthy control's normative performance on the TMT. Z-scores may be converted to T-scores by the following transformation ($T = z * 10 + 50$).

Comparisons of proposed norms to published norms

As the published norms by Heaton et al. (2004) and Tombaugh (2004) are only provided for basic measures (TMT-A and TMT-B), comparisons with the current proposed norms did not include derived measures (TMT B-A, B/A, and β). Proposed norms with only age as a covariate was also not compared to published norms since neither Heaton et al. (2004) nor Tombaugh (2004) offer this option. Normative performance (T-scores) on the TMT measures was calculated for the control group ($n = 292$) following the method described in the previous passage. Next, T-scores were calculated using published norms from Heaton et al. (2004) and Tombaugh (2004). This resulted in three sets of demographically adjusted T-scores, which were compared using paired samples *t*-tests. The control group ($n = 292$) was then split based on the median level of education into a low education group (<13 years of education) and a high education group (≥ 13 years of education) and demographically adjusted T-scores were again compared with paired samples *t*-tests to investigate differences in normative estimations. Distribution of T-scores was assessed with Shapiro–Wilks test of normality and visual comparison with histograms. Norms from Tombaugh (2004) were calculated based on mean scores and standard deviations reported in Tombaugh (2004) and then transformed to T-scores. In some cases, this provided highly abnormal T-scores <0 due to narrow standard deviations in certain stratifications of age and education, and negative T-scores were in these cases set to 0.

Multiple regression analyses were conducted using the same predictors (age, sex, and education) on the *T*-scores derived using norms from Heaton et al. (2004), Tombaugh (2004) and the current proposed Scandinavian norms (Table 2). Reasoning that these *T*-scores should be adjusted for demographic variables (e.g. differences in age should already be corrected for), we expect that results will not be statistically significant ($p \geq 0.05$) if *T*-scores adequately adjust for the demographical variables. Significant effects of any predictor variable would suggest that norms did not adequately correct for the demographical variable when applied to the Scandinavian sample.

Lastly, we examined relationships (Pearson's *r*) between basic measures (TMT-A and TMT-B) and derived measures (TMT B-A and TMT B/A) to the new proposed derived measure TMT- β . All analyses were conducted using the Statistical Package for Social Sciences (SPSS) version 25 and RStudio version 1.2.5033.

Norm calculator

To facilitate the usability and adoption of the proposed regression norms in the clinic, we provide a free web-based tool that computes the regression equations. To obtain normative *T*-scores for both basic (TMT-A and TMT-B) and derived measures (TMT B-A, TMT B/A and TMT- β), the user simply needs to enter valid demographic values (age and years of education) and raw-scores from TMT-A and TMT-B. Except for the TMT B/A, *T*-score calculations are provided for both demographically adjusted norms (age and education) as well as age adjustment only. The tool is implemented as a self-contained HTML/Javascript webpage, available at <https://uit.no/ressurs/uit/cerad/tmt-calc.html> and is released as open source at <https://github.com/DDI-NO/tmt-calc> under Apache License, version 2.0.

Ethics

The Norwegian Regional committees for medical and health research ethics (REK) approved the DDI project from which the current study draws upon. Guidelines in Helsinki declaration of 1964; revised 2013 and the Norwegian Health and Research Act were followed. The Gothenburg MCI study was approved by the local ethics committee and conducted in accordance with the Helsinki declaration. All participants gave written informed consents, including right to withdraw and potential risks and rewards involved.

Results

Effects of demographics on TMT test performance in the healthy control group

Normative regression models and explained variance from predictors for basic and derived measures of the TMT are reported in Table 3. In the following section, improved performance refers to higher scaled scores (Table 2), that is, faster time to completion on basic measures (TMT-A and TMT-B) and reduced difference scores on derived measures (TMT B-A and TMT B/A). On the proposed measure TMT- β , improved performance refers to higher scaled scores on TMT-B adjusting for TMT-A scores.

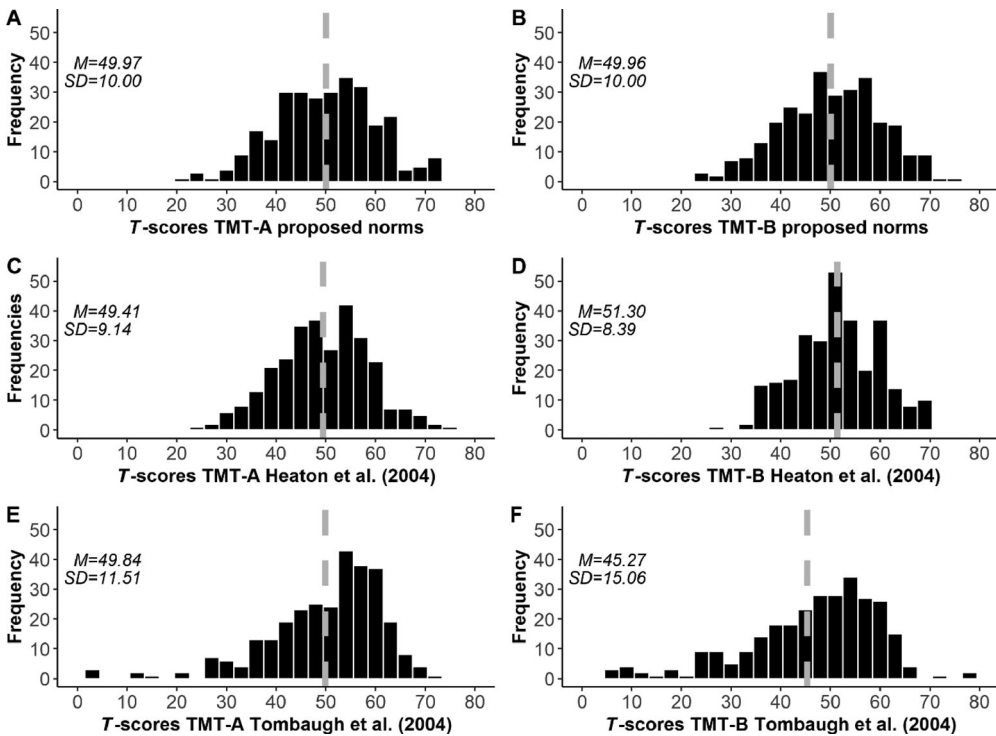


Figure 1. T-score distributions on TMT-A and TMT-B calculated using current proposed norms (A and B) and norms from Heaton et al. (2004) (C and D) and Tombaugh (2004) (E and F) in same control group ($n = 292$). The gray dashed line in each figure depicts the mean T-score for each norm. M and SD are mean and standard deviation, respectively.

Lower age and higher education predicted improved performance on TMT-B, TMT B-A, and TMT- β . On TMT-A, lower age was the only significant predictor for improved performance. Higher education was associated with improved performance on TMT-B, TMT B-A, TMT B/A, and TMT- β . Faster time to completion on TMT-A was associated with improved performance on the proposed measure TMT- β . When omitting education from the normative regression models explained variance from age increased slightly, but total explained variance from the model decreased on all measures.

Adjustment of demographics using published norms

Heaton et al. (2004) norms adequately adjusted for age ($b = -0.103$, $p = 0.101$) on TMT-A. However, Heaton et al. (2004) norms did not adequately correct for the effects of education ($b = -0.771$, partial $r^2 = 0.079$, $p < 0.001$; adjusted $r^2 = 0.076$, $F(3,288) = 8.950$, $p < 0.001$). A similar result was obtained for TMT-B where these norms did not adequately correct for effects of education ($b = -0.661$, partial $r^2 = 0.068$, $p < 0.001$), but adequately adjusted for the effects of age ($b = 0.016$, $p = 0.783$; adjusted $r^2 = 0.062$, $F(3,288) = 7.425$, $p < 0.001$). In contrast, norms from Tombaugh (2004) adequately adjusted for demographics on both TMT-A (adjusted $r^2 = -0.001$, $F(3,288) = 0.877$, $p = 0.453$) and TMT-B (adjusted $r^2 = 0.006$, $F(3,288) = 1.548$, $p = 0.202$).

Table 4. Comparison between normative estimates on the TMT in healthy controls ($n = 292$).

Variable	Test norms	<i>M</i> (<i>SD</i>)	<i>t</i>	<i>df</i>	<i>p</i>	<i>Mdiff</i>	95% CI	
							Lower	Upper
TMT-A	Scandinavian	49.96 (10.00)						
	Tombaugh (2004)	49.86 (11.51)	0.42	291	0.676	0.12	-0.46	0.70
TMT-B	Heaton et al. (2004)	49.41 (9.14)	2.25	291	0.025	0.56	0.07	1.04
	Scandinavian	49.97 (10.00)						
TMT-A < 13 edu	Tombaugh (2004)	45.27 (15.06)	9.26	291	<0.001	4.70	3.70	5.70
	Heaton et al. (2004)	51.30 (8.39)	-5.29	291	<0.001	1.33	-1.83	-0.84
TMT-A ≥ 13 edu	Scandinavian	49.74 (10.15)						
	Tombaugh (2004)	50.37 (11.15)	1.54	142	0.125	0.62	-0.18	1.42
TMT-B < 13 edu	Heaton et al. (2004)	51.85 (9.08)	6.74	142	<0.001	2.10	1.49	2.72
	Scandinavian	50.07 (10.05)						
TMT-A ≥ 13 edu	Tombaugh (2004)	49.30 (12.20)	-1.62	123	0.108	-0.77	-1.72	0.17
	Heaton et al. (2004)	46.95 (8.77)	-12.32	123	<0.001	-3.12	-3.62	-2.62
TMT-B < 13 edu	Scandinavian	49.60 (10.08)						
	Tombaugh (2004)	45.64 (14.61)	-5.56	142	<0.001	-3.96	-5.37	-2.55
TMT-B ≥ 13 edu	Heaton et al. (2004)	53.18 (7.90)	9.78	142	<0.001	3.58	2.86	4.30
	Scandinavian	49.99 (9.97)						
TMT-B < 13 edu	Tombaugh (2004)	45.32 (15.40)	-6.04	123	<0.001	-4.67	-6.20	-3.14
	Heaton et al. (2004)	49.35 (8.68)	-2.41	123	0.017	-0.65	-1.18	-0.12

TMT Scores are *T*-scores adjusted for pertinent demographics. Tombaugh (2004) and Heaton et al. (2004) *T*-scores always compared to Scandinavian norms. *t*, the *t*-test statistic; *M*, mean; *SD*, standard deviation; *df*, degrees of freedom; *Mdiff*, mean difference; 95% CI, lower and upper confidence interval of the mean; *p*, *p*-value.

Distributions of *T*-scores using different norms

Visually comparing distributions of *T*-scores on TMT-A and TMT-B (Figure 1) showed differences in expected normal distributions. Distributions were normal and approximately similar between Heaton et al. (2004) and current proposed norms on TMT-A and TMT-B. In contrast, *T*-scores calculated using Tombaugh (2004) norms showed a non-normal distribution on TMT-A ($W(292) = 0.919$, $p < 0.001$) with a negative skew (-1.270) and leptokurtic kurtosis ($kurtosis = 2.428$). In addition, visually comparing the distribution showed a marked negative tail indicating an increased number of abnormal *T*-scores. This was also observed on TMT-B with Tombaugh (2004) *T*-scores ($skew = -1.11$, $kurtosis = 1.27$, $W(292) = 0.916$, $p < 0.001$).

Comparisons between mean normative estimates

Table 4 compares mean *T*-scores applying norms from Heaton et al. (2004) and Tombaugh (2004) with current proposed norms. On TMT-A, Tombaugh (2004) norms were not significantly different, but Heaton et al. (2004) norms produced lower mean *T*-scores. On TMT-B, Heaton et al. (2004) norms estimated higher mean *T*-scores and Tombaugh (2004) estimated considerably lower scores on TMT-B. Splitting the sample based on educational level showed that for individuals with less than 13 years of education, Heaton et al. (2004) norms produced higher *T*-scores and conversely produced lower *T*-scores for individuals with 13 or more years of education.

Correlations between TMT-β, TMT-A, TMT-B, B-A, and B/A

Correlations between all TMT *T*-score measures are shown in Table 5. A strong association was found between TMT-β and derived measure TMT B-A sharing 93.9% of the variance between measures. Both TMT B-A and TMT-β were highly correlated with

Table 5. Correlations between *T*-scores applying current proposed norms ($n = 292$).

TMT measures	TMT-A	TMT-B	TMT B-A	TMT- β
TMT-A	–			
TMT-B	0.457*	–		
TMT B-A	0.000 ^{ns}	0.864*	–	
TMT- β	–0.003 ^{ns}	0.886*	0.969*	–
TMT B/A	–0.522*	0.492*	0.830*	0.823*

* <0.001 .^{ns}, non-significant result.

TMT-B sharing 74.7% and 78.5% of the variance, respectively. Both TMT B-A and TMT- β were associated with the TMT B/A measure sharing 68.9% and 67.7% of the variance, respectively. Neither TMT B-A nor TMT- β were associated with TMT-A indicating that performance on TMT-A had been adjusted for in both measures.

Discussion

In this study, we propose demographically adjusted test norms for basic and derived measures of TMT in a sample of Scandinavian adults between 41 and 84 years. We compared the proposed test norms to published norms from Heaton et al. (2004) and Tombaugh (2004) and assessed if these norms adequately adjust for demographics when applied to a Scandinavian sample. In addition, we propose a new regression-based approach for estimating the derived TMT B-A measure named TMT- β .

The effects of age on TMT-A and TMT-B were comparable to other regression-based norms with a similar age demographic (Peña-Casanova et al., 2009). Conversely, education accounted for much less variance on TMT-A and TMT-B (Gonçalves et al., 2013; Peña-Casanova et al., 2009). This discrepancy is likely due to differences in sample composition between the examined studies. For instance, Peña-Casanova et al. (2009) reported that about 20% of participants attained ≤ 5 years of education and over 20% attained ≥ 16 years. In contrast, the normative sample of Scandinavians employed in the current study had no participants with less than 6 years of education and generally a high level of education ($M = 13.21$, $SD = 3.34$; Table 1). Thus, the normative sample of Scandinavians had a restricted range of education compared to Peña-Casanova et al. (2009) which might explain why education accounted for less variance. While we believe the educational level observed in the Scandinavian sample is representative of the Scandinavian population (Eurostat, 2019), homogenic high levels of education limits the applicability of the norms to countries with a similar educational composition. Discrepancy between demographics of the initial normative sample and the target population where the norms are applied must be considered for reliable normative estimation, as argued by Heaton et al. (1999). Finally, sex did not contribute significantly to scores on any TMT measure which is consistent with most normative studies (Mitrushina et al., 2005, p. 69). As expected from earlier studies, derived measures TMT B-A and TMT B/A were less influenced by age and education than basic measures (Bezdicsek et al., 2012; Gonçalves et al., 2013; Hester et al., 2005; Periañez et al., 2007; Sanchez-Cubillo et al., 2009). As a result, adjusting for demographics on derived measures has less impact on normative estimations, but appropriate normative data based on a representative sample should still be used for reliable estimations.

Derived measures of TMT are employed to minimize the impact of visual search/visuomotor demands and subsequently enhance measurement of executive functioning associated with TMT-B. As an alternative approach to TMT B-A, we reasoned that we could regress TMT-A scores alongside pertinent demographics on TMT-B scores which would isolate the higher order executive functions associated with TMT-B. This new measure was named TMT- β to avoid confusion with the conventional TMT B-A approach. While the demographically adjusted TMT B-A and TMT- β *T*-scores were highly correlated in our sample (93.9% shared variance), TMT- β might still provide utility in clinical samples where both TMT-A and TMT-B is slow due to visual scanning and/or visuomotor deficits. This would result in an elevated difference score TMT B-A, thus giving the appearance of executive function deficits. Senior et al. (2018) showed that slow time to completion on both TMT-A and TMT-B occurred in 37% of cases in a clinical sample but when compared to others with similar TMT-A scores, 40% of these did not show a disproportionate increase in TMT-B, indicating that executive deficits were not the primary cause of the abnormal TMT B-A difference. Compared to the conventional TMT B-A measure, TMT- β should in these instances be able to discern the individuals who do not show a disproportionate increase in TMT-B completion times by adjusting scores based on their individual TMT-A completion time. As an example, a 75-year-old individual from a clinical sample with 9 years of education completing TMT-A in 71 s and TMT-B in 202 s estimates a demographically adjusted *T*-score of 25 on TMT B-A applying current proposed norms. In contrast, the same individual would receive a *T*-score of 35 on TMT- β . This indicates that TMT B-A may produce disproportionately low estimates of executive function as compared to the TMT- β when both TMT-A and TMT-B completion times are slow. TMT- β differs from the stratified approach used by Senior et al. (2018) as we employ multiple regression analysis to adjust for TMT-A completion time. This allows for the adjustment of TMT-A performance at a continuous level while at the same time correcting for normative effects of age and education. We have introduced TMT- β with some potential advantages discussed, but further research into criterion validity and clinical applications need to be established. Compared with the traditional TMT B-A measurement, we hypothesize that TMT- β should be better able to discern individuals with abnormal TMT B-A scores, and therefore correlate more strongly with cognitive flexibility and associated brain structures, particularly in clinical samples.

A key objective of this study was to compare norms from Heaton et al. (2004), Tombaugh (2004) and the current proposed norms in a Scandinavian sample. While the Heaton et al. (2004) norms produced apparently similar distributions of *T*-scores as current proposed norms (Figure 1), results from multiple regression analysis showed that significant effects of education were still evident on TMT-A (7.8%) and TMT-B (6.8%). The associated beta coefficients were negative, suggesting that the Heaton et al. (2004) norms generally overestimated the significance of education when applied in the Scandinavian sample. Individuals with lower educational attainment had significantly higher *T*-scores than expected while individuals with higher educational attainment had lower *T*-scores (Table 4). On TMT-A, Heaton et al. (2004) reported 10% explained variance from education, however no effects of education were evident in the Scandinavian sample on TMT-A. On TMT-B, Heaton et al. (2004) reported 16% on

education compared with 4% in the Scandinavian sample. Thus, education accounted for larger amounts of variability in the initial normative sample employed in the Heaton et al. (2004) norms, providing a likely explanation for why norms overestimated the effects of education when applied in the Scandinavian sample. Education is generally considered more affordable and available to the public in Scandinavian countries which might be why education apparently has less impact on scores. Future normative studies in Scandinavia should compare the effects of demographic corrections to investigate if this applies to other neuropsychological measures as well.

T-scores from Tombaugh (2004) produced non-normal distributions with a negative skew and leptokurtic kurtosis (Figure 1) and subsequently lower mean scores on TMT-B (Table 4). This likely stems from narrow standard deviations of mean scores in certain stratifications of age and education in the Tombaugh (2004) sample, whereby slight deviation in scores result in highly abnormal *T*-scores for a substantial proportion of the Scandinavian sample. In terms of demographic corrections, however, results from multiple regression analysis showed that *T*-scores from Tombaugh (2004) adequately adjusted for age and education in the Scandinavian sample. Tombaugh (2004) also reported that age was the largest contributor to variance on TMT-A and TMT-B with only marginal effects of education. Results from multiple regression analysis suggested that normative estimates were comparable to the Scandinavian sample.

We provide normative regression models omitting education as a covariate. Education may not always be a relevant normative demographic for all target populations (e.g. low educational attainment while scoring above average on age adjusted measures of intelligence). The implications of using these norms for individuals with low educational attainment are slightly stricter normative corrections (i.e. lower *T*-scores). It can be appropriate to use these norms in instances where an individual did not have the opportunity for education that they otherwise would have benefited from. These norms should not be applied to individuals who lack education because they could not comprehend the material or otherwise were not eligible (Mitrushina et al., 2005, p. 31). Our results indicate that age accounted for slightly more variance in scores when omitting education as a covariate but overall explained variance in the models decreased (Table 3). Norms correcting for all pertinent demographics should therefore be used when appropriate.

Some limitations need to be addressed. First, an important limitation of this study was the lack of an independent sample of healthy controls to apply and assess our proposed norms. We therefore opted to compare current proposed norms to published norms within in the same sample ($n=292$). Second, healthy controls enrolled in the normative sample were not screened for perceptual-motor deficits which might inhibit performance on the TMT prior to testing. Lastly, it is important to emphasize that the current proposed norms are not *better* than the published norms, but simply that there is an advantage to applying local norms, as shown when comparing current proposed norms to published norms in the Scandinavian sample. We also stress that the users of the current proposed norms should follow the same administration procedures on TMT for reliable estimates, which are described in Strauss et al. (2006).

Conclusions

We propose demographically adjusted regression-based norms for age 41 through 84 years on TMT-A and TMT-B and derived measures TMT B-A and TMT B/A based on healthy controls from the Norwegian DDI and Swedish Gothenburg MCI cohorts. We also propose a new measure named TMT- β developed using a regression-based procedure to improve on the conventional TMT B-A. Comparisons of norms from Heaton et al. (2004) and Tombaugh (2004) suggest that current proposed norms are better suited for use in a Scandinavian population. To ease the use and availability of the regression norms in clinical settings, a free online norm calculator is offered <https://uit.no/ressurs/uit/cerad/tmt-calc.html>.

Acknowledgments

We thank Svein Ivar Bekkelund, Kjell-Arne Arntzen, Kai Müller, Claus Albretsen, Mari Thoresen Løkholm, Ida Harviken, Line Saether, Ingrid Myrvoll Lorentzen, Erna Utnes, Marianne Wettergreen, Berglind Gisladdottir, Marit Knapstad, Reidun Meling, Synnøve Bremer Skarpenes and Elin Margrethe Solli for clinical examinations and essential help with the project.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy of the research participants.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the University of Tromsø—the Arctic University of Norway; the Norwegian Research Council (Dementia Disease Initiation) under Grant number 217780; Helse Sør-øst, NASATS (Dementia Disease Initiation) under Grant number 2013131 and Helse Nord under grant number HNF1401-18. Additional support was received from the Sahlgrenska University Hospital, the Swedish Research Council, Swedish Brain Power, the Swedish Dementia Foundation, the Swedish Alzheimer Foundation, Stiftelsen Psykiatriska forskningsfonden, and Konung Gustaf V:s and Drottning Victorias Frimurarestiftelse. The funding sources were not involved in the drafting of this manuscript.

ORCID

Jacob Espenes  <http://orcid.org/0000-0002-2383-5348>
Ingvild Vøllo Eliassen  <http://orcid.org/0000-0003-1288-6032>
Knut Waterloo  <http://orcid.org/0000-0003-3447-8312>
Tormod Fladby  <http://orcid.org/0000-0002-9984-9797>
Bjørn-Eivind Kirsebom  <http://orcid.org/0000-0002-1413-9578>

References

- Abi Chahine, J., Rammal, S., Fares, Y., & Abou Abbas, L. (2019). Trail making test: Normative data for the Lebanese adult population. *The Clinical Neuropsychologist*, 4. <https://doi.org/10.1080/13854046.2019.1701710>
- Arbuthnott, K., & Frank, J. (2000). Executive control in set switching: residual switch cost and task-set inhibition. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Experimentale*, 54(1), 33–41. <https://doi.org/10.1037/h0087328>
- Bezdicek, O., Motak, L., Axelrod, B. N., Preiss, M., Nikolai, T., Vyhnalek, M., Poreh, A., & Ruzicka, E. (2012). Czech version of the trail making test: normative data and clinical utility. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 27(8), 906–914. <https://doi.org/10.1093/arclin/acs084>
- Cavaco, S., Gonçalves, A., Pinto, C., Almeida, E., Gomes, F., Moreira, I., Fernandes, J., & Teixeira-Pinto, A. (2013). Trail making test: regression-based norms for the Portuguese Population. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 28(2), 189–198. <https://doi.org/10.1093/arclin/acs115>
- Dickinson, M. D., & Hiscock, M. (2011). The Flynn effect in neuropsychological assessment. *Applied Neuropsychology*, 18(2), 136–142. <https://doi.org/10.1080/09084282.2010.547785>
- Dodge, H. H., Zhu, J., Lee, C.-W., Chang, C.-C H., & Ganguli, M. (2014). Cohort effects in age-associated cognitive trajectories. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, 69(6), 687–694. Retrieved from doi:<https://doi.org/10.1093/gerona/glt181>
- Eurostat (2019). *Population by educational attainment level, sex and age (%)*. Retrieved from: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=edat_ifs_9903&lang=en&fbclid=IwAR1fKmi-U_BrKWdSr6v5yCpTAv3LgaJuLBS4IDnZqV2LLvKXhXvihB1WWRM
- Fernandez, A. L., & Marcolulos, B. A. (2008). A comparison of normative data for the trail making test from several countries: equivalence of norms and considerations for interpretation. *Scandinavian Journal of Psychology*, 49(3), 239–246. Retrieved from <https://doi.org/10.1111/j.1467-9450.2008.00637.x>
- Fladby, T., Pålhaugen, L., Selnes, P., Waterloo, K., Bråthen, G., Hessen, E., Almdahl, I. S., Arntzen, K.-A., Auning, E., Eliassen, C. F., Espenes, R., Grambaite, R., Grøntvedt, G. R., Johansen, K. K., Johnsen, S. H., Kalheim, L. F., Kirsebom, B.-E., Müller, K. I., Nakling, A. E., ... Aarsland, D. (2017). Detecting at-risk Alzheimer's disease cases. *Journal of Alzheimer's Disease*, 60(1), 97–105. Retrieved from <https://doi.org/10.3233/JAD-170231>
- 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. <https://doi.org/10.1080/01688639508405143>
- Goul, W. R., & Brown, M. (1970). Effects of age and intelligence on trail making test performance and validity. *Perceptual and Motor Skills*, 30(1), 319–326. <https://doi.org/10.2466/pms.1970.30.1.319>
- Heaton, R. K., Avitable, N., Grant, I., & 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(4), 572–582. <https://doi.org/10.1076/jcen.21.4.572.882>
- Heaton, R., Miller, S., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults*. Psychological Assessment Resources.
- Hester, R. L., Kinsella, G. J., Ong, B., & McGregor, J. (2005). Demographic influences on baseline and derived scores from the trail making test in healthy older Australian adults. *The Clinical Neuropsychologist*, 19(1), 45–54. <https://doi.org/10.1080/13854040490524137>
- Kennedy, K. J. (1981). Age effects on trail making test performance. *Perceptual and Motor Skills*, 52(2), 671–675. <https://doi.org/10.2466/pms.1981.52.2.671>
- Kirsebom, B. E., Espenes, R., Hessen, E., Waterloo, K., Harald Johnsen, S., Gundersen, E., ... Fladby, T. (2019). Demographically adjusted CERAD wordlist test norms in a Norwegian sample from 40 to 80 years. *The Clinical Neuropsychologist*, 33:sup1, 27-39, DOI: <https://doi.org/10.1080/13854046.2019.1574902>

- Lamberty, G. J., Putnam, S. H., Chatel, D. M., & Bieliauskas, L. A. &. (1994). Derived trail making test indices: a preliminary report. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology*, 7(3), 230–234.
- Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). Oxford University Press.
- Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment*. Oxford University Press.
- Moll, J., Oliveira-Souza, R. d., Moll, F. T., Bramati, I. E., & Andreiuolo, P. A. (2002). The cerebral correlates of set-shifting: an fMRI study of the trail making test. *Arquivos de Neuro-Psiquiatria*, 60(4), 900–905. <https://doi.org/10.1590/s0004-282x2002000600002>
- Peña-Casanova, J., Quiñones-Ubeda, S., Quintana-Aparicio, M., Aguilar, M., Badenes, D., Molinuevo, J. L., Torner, L., Robles, A., Barquero, M. S., Villanueva, C., Antúnez, C., Martínez-Parra, C., Frank-García, A., Sanz, A., Fernández, M., Alfonso, V., Sol, J. M., & Blesa, R., NEURONORMA Study Team (2009). Spanish Multicenter Normative Studies (NEURONORMA Project): norms for verbal span, visuospatial span, letter and number sequencing, trail making test, and symbol digit modalities test. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 24(4), 321–341. <https://doi.org/10.1093/arclin/acp038>
- Periáñez, J. A., Ríos-Lago, M., Rodríguez-Sánchez, J. M., Adrover-Roig, D., Sánchez-Cubillo, I., Crespo-Facorro, B., Quemada, J. I., & Barceló, F. (2007). Trail making test in traumatic brain injury, schizophrenia, and normal ageing: sample comparisons and normative data. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 22(4), 433–447. Retrieved from <https://doi.org/10.1016/j.acn.2007.01.022>
- Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation* (Vol. 4): Reitan Neuropsychology.
- Reitan, R. M., & Wolfson, D. (2004). The trail making test as an initial screening procedure for neuropsychological impairment in older children. *Archives of Clinical Neuropsychology*, 19(2), 281–288. [https://doi.org/10.1016/S0887-6177\(03\)00042-8](https://doi.org/10.1016/S0887-6177(03)00042-8)
- Sanchez-Cubillo, I., Perianez, J. A., Adrover-Roig, D., Rodríguez-Sánchez, J. M., Ríos-Lago, M., Tirapu, J., & Barcelo, F. (2009). Construct validity of the trail making test: role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *Journal of the International Neuropsychological Society*, 15(3), 438–450. <https://doi.org/10.1017/s1355617709090626>
- Senior, G., Piovesana, A., & Beaumont, P. (2018). Discrepancy analysis and Australian norms for the Trail Making Test. *The Clinical Neuropsychologist*, 32(3), 510–523. <https://doi.org/10.1080/13854046.2017.1357756>
- Siciliano, M., Chiorri, C., Battini, V., Sant'Elia, V., Altieri, M., Trojano, L., & Santangelo, G. (2019). Regression-based normative data and equivalent scores for trail making test (TMT): an updated Italian normative study. *Neurological Sciences*, 40(3), 469–477. <https://doi.org/10.1007/s10072-018-3673-y>
- St-Hilaire, A., Parent, C., Potvin, O., Bherer, L., Gagnon, J.-F., Joubert, S., Belleville, S., Wilson, M. A., Koski, L., Rouleau, I., Hudon, C., & Macoir, J. (2018). Trail making tests A and B: regression-based normative data for Quebec French-speaking mid and older aged adults. *The Clinical Neuropsychologist*, 32(sup1), 77–90. <https://doi.org/10.1080/13854046.2018.1470675>
- Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. American Chemical Society.
- Strobel, C., Johansen, H., Aga, O., Bekkhus-Wetterberg, P., Brierly, M., Egeland, J., Follesø, K., Rike, P., Schanke, A. (2018). Manual Norsk Revidert trail making test (TMT-NR3). Retrieved from https://aldring-og-helse-media.s3.amazonaws.com/documents/TMT-NR3_AoH_Manual_2018_4UhFcRC.pdf
- Stuss, D. T., Stethem, L. L., & Pelchat, G. (1988). Three tests of attention and rapid information processing: an extension. *Clinical Neuropsychologist*, 2(3), 246–250. <https://doi.org/10.1080/13854048808520107>
- Testa, S. M., Winicki, J. M., Pearlson, G. D., Gordon, B., & Schretlen, D. J. (2009). Accounting for estimated IQ in neuropsychological test performance with regression-based techniques. *Journal of the International Neuropsychological Society*, 15(6), 1012–1022. Retrieved from

<https://www.cambridge.org/core/article/accounting-for-estimated-iq-in-neuropsychological-test-performance-with-regressionbased-techniques/8D4BCB20747A14F10D656971A9F96160>. <https://doi.org/10.1017/S1355617709990713>

Tombaugh, T. N. (2004). Trail making test A and B: normative data stratified by age and education. *Archives of Clinical Neuropsychology*, *19*(2), 203–214. Retrieved from [https://doi.org/10.1016/S0887-6177\(03\)00039-8](https://doi.org/10.1016/S0887-6177(03)00039-8)

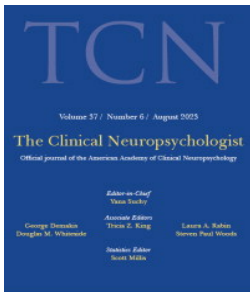
Wallin, A., Nordlund, A., Jonsson, M., Lind, K., Edman, Å., Göthlin, M., Stålhammar, J., Eckerström, M., Kern, S., Börjesson-Hanson, A., Carlsson, M., Olsson, E., Zetterberg, H., Blennow, K., Svensson, J., Öhrfelt, A., Bjerke, M., Rolstad, S., & Eckerström, C. (2016). The Gothenburg MCI study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism*, *36*(1), 114–131. <https://doi.org/10.1038/jcbfm.2015.147>

Paper 2

Espenes, J., Eliassen, I. V., Öhman, F., Hessen, E., Waterloo, K., Eckerström, M., ... & Kirsebom, B.E. (2022).

Regression-based normative data for the Rey Auditory Verbal Learning Test in Norwegian and Swedish adults aged 49–79 and comparison with published norms.

The Clinical Neuropsychologist, 37(6), 1276-1301.



Regression-based normative data for the Rey Auditory Verbal Learning Test in Norwegian and Swedish adults aged 49–79 and comparison with published norms

Jacob Espenes, Ingvild Vøllo Eliassen, Fredrik Öhman, Erik Hessen, Knut Waterloo, Marie Eckerström, Ingrid Myrvoll Lorentzen, Cecilie Bergland, Madelene Halvari Niska, Santiago Timón-Reina, Anders Wallin, Tormod Fladby & Bjørn-Eivind Kirsebom

To cite this article: Jacob Espenes, Ingvild Vøllo Eliassen, Fredrik Öhman, Erik Hessen, Knut Waterloo, Marie Eckerström, Ingrid Myrvoll Lorentzen, Cecilie Bergland, Madelene Halvari Niska, Santiago Timón-Reina, Anders Wallin, Tormod Fladby & Bjørn-Eivind Kirsebom (2023) Regression-based normative data for the Rey Auditory Verbal Learning Test in Norwegian and Swedish adults aged 49–79 and comparison with published norms, *The Clinical Neuropsychologist*, 37:6, 1276-1301, DOI: [10.1080/13854046.2022.2106890](https://doi.org/10.1080/13854046.2022.2106890)

To link to this article: <https://doi.org/10.1080/13854046.2022.2106890>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 13 Aug 2022.



[Submit your article to this journal](#)



Article views: 1538



[View related articles](#)










[View Crossmark data](#)



Citing articles: 2 [View citing articles](#)

Regression-based normative data for the Rey Auditory Verbal Learning Test in Norwegian and Swedish adults aged 49–79 and comparison with published norms

Jacob Espenes^{a,b} , Ingvild Vøllo Eliassen^{c,d} , Fredrik Öhman^{e,f} , Erik Hessen^{c,d}, Knut Waterloo^{a,b} , Marie Eckerström^e, Ingrid Myrvoll Lorentzen^a , Cecilie Bergland^a, Madelene Halvari Niska^a, Santiago Timón-Reina^{c,g}, Anders Wallin^e, Tormod Fladby^{c,h}  and Bjørn-Eivind Kirsebom^{a,b} 

^aDepartment of Psychology, Faculty of Health Sciences, The Arctic University of Norway, Tromsø, Norway; ^bDepartment of Neurology, University Hospital of North Norway, Tromsø, Norway; ^cDepartment of Neurology, Akershus University Hospital, Lørenskog, Norway; ^dDepartment of Psychology, University of Oslo, Oslo, Norway; ^eDepartment of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; ^fWallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Gothenburg, Sweden; ^gDepartamento de Inteligencia Artificial, Universidad Nacional de Educación a Distancia, Madrid, Spain; ^hInstitute of Clinical Medicine, University of Oslo, Oslo, Norway

ABSTRACT

Objective: The Rey Auditory Verbal Learning Test (RAVLT) is a widely used measure of episodic verbal memory. To our knowledge, culturally adapted and demographically adjusted norms for the RAVLT are currently not available for Norwegian and Swedish adults, and imported North American norms are often used. We here develop regression-based norms for Norwegian and Swedish adults and compare our norms to North American norms in an independent sample of cognitively healthy adults. **Method:** Participants were 244 healthy adults from Norway and Sweden between the aged 49 and 79 years, with between 6 and 24 years of education. Using a multiple multivariate regression-based norming procedure, we estimated effects of age, sex, and years of education on basic and derived RAVLT test scores. The newly developed norms were assessed in an independent comparison group of cognitively healthy adults ($n=145$) and compared to recently published North American regression-based norms. **Results:** Lower age, female sex and more years of education predicted higher performance on the RAVLT. The new norms adequately adjusted for age, education, and sex in the independent comparison group. The American norms corrected for demographics on all RAVLT trials except trials 4, 7, list B, and trials 1–5 total. Test-retest ($M=2.55$ years) reliability varied from poor to good. **Conclusion:** We propose regression-based norms for the RAVLT adjusting for pertinent demographics. The norms may be used for assessment of Norwegian and Swedish adults between the aged of 49 and 79 years, with between 6 and 24 years of education.

ARTICLE HISTORY

Received 25 March 2022
Accepted 23 July 2022
Published online 16 August 2022

KEYWORDS

Normative;
Rey Auditory Verbal Learning Test;
Norway; Sweden;
memory

CONTACT Jacob Espenes  Johan.j.espenes@uit.no  Department of Psychology, Faculty of Health Sciences, The Arctic University of Norway, Hansine Hansens veg 18, Tromsø, 9037, Norway.

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Introduction

The Rey Auditory Verbal Learning Test (RAVLT) is a widely used measure of episodic verbal memory in the field of neuropsychology (Boake, 2000). It is a multi-trial, 15-item word list test that enables assessment of fundamental memory processes, including acquisition, interference effects, retention, and retrieval (Ivnik et al., 1992). The RAVLT is sensitive to learning and memory deficits in several clinical groups, including patients with mild cognitive impairment (MCI; Estévez-González et al., 2003), Alzheimer's Disease (AD; Ricci et al., 2012), left hemispheric brain pathology (Loring et al., 2008), and neuropathologies of various etiologies (Powell et al., 1991). RAVLT scores are good markers for progressive episodic memory deficits typical in common age-related conditions such as AD and MCI due to AD (Belleville et al., 2017). Delayed recall performance on the RAVLT has demonstrated adequate to excellent diagnostic accuracy for identifying which individuals with MCI will progress to AD dementia (Eckerström et al., 2013; Ewers et al., 2012).

Sociodemographic factors have been found to influence RAVLT performance. Age effects are consistently reported in middle-aged and older adults, showing declining performance with increasing age (Lavoie et al., 2018; Messinis et al., 2016; Stricker et al., 2021). Findings are somewhat less consistent regarding the influence of sex and educational attainment. Several studies indicate a clear female advantage on RAVLT performance (Asperholm et al., 2019; Lavoie et al., 2018; Stricker et al., 2021; Sundermann et al., 2016; 2017; Van Der Elst et al., 2005) while others find no significant influence of sex (Marqués et al., 2013; Messinis et al., 2016). In contrast, an older meta-analytic review of demographic influences on RAVLT performance suggests a male advantage on some trials, and otherwise no effects of sex on performance (Mitrushina et al., 2005). Individuals with more years of education often obtain higher scores on the RAVLT (Bezdicek et al., 2014; Lavoie et al., 2018; Messinis et al., 2016; Stricker et al., 2021; Van Der Elst et al., 2005). However, meta-analytic evidence has indicated no significant effect of education on performance (Mitrushina et al., 2005).

Linguistic and cultural differences may also contribute to systematic variation of RAVLT performance in different populations. Norms from different cultural groups are not necessarily interchangeable. Which norm set we choose to apply on an individual's scores may influence their likelihood of being classified as memory-impaired (Strauss et al., 2006). Local norms for the RAVLT have been developed for older adults with diverse cultural and linguistic backgrounds including north American (Stricker et al., 2021), Venezuelan (Correia & Osorio, 2014), French-Canadian (Lavoie et al., 2018), Greek (Messinis et al., 2016), Israeli (Vakil et al., 2010), and German (Boenniger et al., 2021).

To our knowledge, there are currently no demographically adjusted test norms for the RAVLT available for the Norwegian or Swedish middle-aged and older adults. Demographically adjusted and locally sourced normative material is needed to increase the likelihood for an accurate evaluation of memory function in this population. Thus, the first objective of this study was to develop normative data for the RAVLT for Norwegian and Swedish adults ages 49 to 79 years applying a multiple multivariate approach (Van der Elst et al., 2017). Secondly, clinicians in Norway and Sweden have several sets of norms available for use such as the newly developed

population-based regression-based norms from the Mayo Normative Study (Stricker et al., 2021) that may or may not be appropriate in a Scandinavian population. Thus, the study's second objective was to compare the currently proposed norms with published norms from Stricker et al. (2021) in an independent sample of cognitively healthy participants with subjective cognitive decline (SCD) from Norway and Sweden.

Methods and materials

Participants

The present study included 244 healthy control participants from three related research projects on early phases of dementia diseases conducted in Norway and Sweden; the Dementia Initiation study (DDI, $n=70$); the Gothenburg Mild Cognitive Impairment (MCI) study ($n=121$); and the Oslo MCI study ($n=53$). Healthy controls included from DDI were assessed at the Akershus University Hospital or the University Hospital of Northern Norway between January 2013 and June 2020. The Oslo MCI study is the predecessor of the ongoing DDI study, and assessments were performed at the Akershus University hospital between 2005 and 2013. Participants included from the Gothenburg MCI study were assessed at the Sahlgrenska University Hospital, Sweden, between January 2001 and March 2014. Healthy controls from DDI- ($n=70$) and the Oslo MCI- study ($n=70$) were primarily recruited from spouses of symptom group participants and secondarily through advertisements in local media and from the orthopedic wards. Healthy controls included from the Gothenburg MCI study ($n=121$) were primarily recruited through senior citizen organizations, and a small proportion were relatives of symptom group participants. All studies followed a similar standardized procedure for assessment that included neurological and physical examination, neuropsychological assessment and self and informant-reported medical history. Most participants agreed to submit blood samples and cerebrospinal fluid samples. However, these were not analyzed for the purpose of this study. For a complete description of the Gothenburg MCI cohort, methods, and study procedures, see Wallin et al. (2016). For DDI see Fladby et al. (2017) and for Oslo MCI refer to Hessen et al. (2014).

Joint criteria for inclusion applied to all healthy controls employed in the normative analyses of the present study ($n=244$) was aged 49 through 79, the absence of subjective symptoms of cognitive decline, mini mental state examination (MMSE) ≥ 26 , and a native language of Norwegian or Swedish. The normative sample was split between 122 participants speaking Norwegian and 122 speaking Swedish. Two participants spoke Norwegian as the second language. Fifty participants were between aged 49 and 58; 122 were between 59 and 68 years; and 72 were between 69 and 79 years. Education ranged between 6 and 24 years of education. Every full year of formal education attained by the participants was counted, excluding degrees of the same level. Exclusion criteria were developmental disorders, neurological disease, intellectual disability, severe somatic disorders that might negatively influence cognitive performance, history of stroke, or severe psychiatric disorder, including major depression. Apart from MMSE, results on cognitive screening tests and neuropsychological measures were not used to verify cognitive normalcy or exclude participants as this potentially excludes normal healthy participants, thereby reducing variation

associated with normal aging, thus limiting the generalizability and validity of the norms. Scores from participants who did not complete the RAVLT or had missing scores on any RAVLT trial was excluded from the analysis. Thus, only participants with complete RAVLT administrations were included.

Between cohort comparisons of demographics and cognitive performance

Participants were recruited from three related research projects, and potential cohort effects were investigated. While the Gothenburg and -Oslo MCI cohort participants on average had fewer years of education, there were no cohort effects on RAVLT raw scores adjusted for age differences, years of education, and sex, except for trial 1. Scores on trial 1 were analyzed in a regression model, which included the predictor's cohort (dummy coded to account for three cohorts), age, years of education, and sex. Results showed that control participants recruited from the Oslo MCI study on average remembered 0.738 fewer words than the controls from DDI, adjusting for differences in education, age and sex ($b = -0.738$, 95% CI $[-1.327, -.150]$, $p = .014$, $F(5, 238) = 10.246$, and $p = <.001$). There were no significant differences between the Oslo MCI cohort and participants from the Gothenburg MCI study.

Independent comparison group to assess norms

The DDI study and Gothenburg MCI study also include participants with subjective cognitive decline (SCD), and at the time of analysis, 145 cognitively healthy participants with SCD had available assessments on the RAVLT. These were included in a separate sample to evaluate and compare the current proposed norms with published norms from Stricker et al. (2021). All SCD participants underwent the same standardized procedure for assessment as previously described for healthy controls, including the general exclusion and inclusion criteria, and MMSE score ≥ 26 . SCD participants were included via referrals from general practitioners to memory clinics, and self-referral following public advertisements aimed at individuals with memory complaints. As such, memory deficits were the main cognitive complaints. SCD was determined by self-report the following proposed guidelines in Jessen et al. (2014) and Molinuevo et al. (2017). All participants with SCD were subject to a clinical interview about the nature of progression since onset, experience of cognitive deficits in other domains, familiar history, and affective symptoms. To ensure cognitive normalcy and differentiate participants presenting SCD from MCI, recommendations from Albert et al. (2011) were applied and participants were excluded if they presented *objective* cognitive decline, operationalized as a score 1.5 standard deviation below the normative mean on at least one of the following neuropsychological tests (applied normative corrections in parenthesis); The Trail Making Test B (Espenes et al., 2020; Reitan & Wolfson, 1985), Controlled Oral Word Association test (COWAT, Heaton et al., 2004; Lorentzen et al., 2021), Silhouettes from Visual Object and Space Perception Battery (VOSP, Eliassen et al., 2020; Warrington & James, 1991). Participants with SCD from the DDI cohort were excluded on basis of the CERAD word list-delayed recall (Fillenbaum et al., 2008; Kirsebom et al., 2019). Participants from the Gothenburg MCI cohort did

not perform the CERAD word list delayed recall and were instead excluded based on the RAVLT trial 7 (Rey, 1958; Stricker et al., 2021).

To investigate if the SCD group would be suitable as an independent group for comparing normative adjustments, their RAVLT scores were compared to those of the controls (Table 1). Regression analyses indicated no significant differences between groups adjusting for differences in years of education, age, and sex, except for trial 4, where a minor difference was observed. The SCD group on average remembered 0.742 more words compared to the healthy controls ($b=0.742$, $p = .032$, and 95% CI [0.063, 1.420]). The confidence interval suggests that this difference could be very small, possibly spurious, as there is no theoretical basis for trial 4 differing substantially from other parts of the RAVLT. We therefore conclude that the comparison group comprised of individuals presenting SCD had comparable scores to the healthy control group on the RAVLT, indicating that they were suitable as an independent comparison group.

RAVLT test version and administration

RAVLT assessments were performed by clinical psychologists or psychologists-in-training. Firstly, the participant is instructed to try to remember as many words as possible from a list of words that is about to be read aloud. Then, a list of 15 words (list A), is read aloud to the participant, to which the participant is required to recall as many words as possible directly after. This is repeated for a total of five trials, and the participant is required to freely recall as many words as possible after each presentation.

Table 1. Demographics, raw scores of the normative sample of healthy controls and the independent comparison group comprised of cognitively healthy participants with subjective cognitive decline (SCD).

Variables	Normative sample of healthy controls ($n=244$)	Independent comparison group ($n=145$)	t^a/χ^2	p
Age Mean (SD) [range]	64.3 (6.8) [49–79]	62.3 (6.7) [49–77]	2.952	.003
Female n (%)	138 (56.6 %)	91 (62.8 %)	1.444	<i>n.s.</i>
Years of education Mean (SD) [range]	12.7 (3.3) [6–24]	14.0 (3.2) [6–21]	–3.666	<.001
Trial 1 Mean (SD)	5.5 (1.7)	5.8 (1.9)	0.260	<i>n.s.</i>
Trial 2 Mean (SD)	8.3 (2.1)	8.7 (2.2)	–0.094	<i>n.s.</i>
Trial 3 Mean (SD)	9.7 (2.4)	10.5 (2.5)	1.306	<i>n.s.</i>
Trial 4 Mean (SD)	10.8 (2.4)	12.0 (4.5)	2.150	.032
Trial 5 Mean (SD)	11.4 (2.5)	12.0 (2.1)	0.397	<i>n.s.</i>
Trial 6 (immediate memory) Mean (SD)	9.3 (3.1)	10.3 (2.6)	1.263	<i>n.s.</i>
Trial 7 (delayed memory) Mean (SD)	9.0 (3.1)	10.1 (2.5)	1.522	<i>n.s.</i>
Trials A1–A5 total Mean (SD)	45.6 (9.6)	48.4 (9.2)	0.687	<i>n.s.</i>
List B Mean (SD)	5.4 (1.8)	5.6 (2.0)	–0.425	<i>n.s.</i>

Notes. n = Number of participants; p = p -value; t = t statistic; *n.s.* = non-significant result ($p > .05$); Results are presented as mean (Standard deviation) [range] except for sex which is characterized by female percentage.

^afor RAVLT scores, test statistics refer to mean difference between groups controlling for age, years of education, and sex. For age and years of education, independent samples t -tests with Welch correction were conducted; χ^2 = Chi Square test for 2×2 table.

After five consecutive trials, a distractor list (B) containing 15 separate words is presented, and the participant is asked to freely recall as many words as possible from this new list. Following immediately, without cues or renewed presentation, the participant is asked to recall list A again (trial 6). After a timed delay of 30 minutes, during which other neuropsychological tests with non-verbal stimuli were conducted, the participant is required to freely recall List A once more (trial 7), reflecting delayed verbal memory. On the RAVLT, the primary variables are correctly recalled words on learning trials (trial 1 to 5), list B, trial 6, and trial 7. In addition, derived measures (Table 2) are often computed to provide evaluations of learning (Ivnik et al., 1990), inhibition and interference effects, and retention (i.e., correctly recalled words after 30-minute delay relative to the number of words previously recalled).

Norwegian participants were administered a Norwegian translation of the RAVLT word list, available in English in Lezak et al. (2012). Likewise, the Swedish participants from the Gothenburg MCI cohort used a translated Swedish version. English, Norwegian, and Swedish versions of the RAVLT with standardized instructions and word lists A and B are presented in appendix A. Norwegian and Swedish versions of the RAVLT were not backtranslated or otherwise formally validated. The Swedish Gothenburg MCI study employed a different protocol for administering the recognition trial than the Norwegian cohorts from DDI and Gothenburg-Oslo MCI and we therefore do not present normative data for this part of the test.

Regression norming procedure

Following procedures described in Van der Elst et al. (2017), multivariate regression-based norms were developed based on the performance of the included healthy controls ($n=244$) on all primary RAVLT measures. Exploratory analyses confirmed that all primary RAVLT measures were moderately to highly correlated ($r = .289-.868$), suggesting that

Table 2. Primary and the derived measures on the RAVLT.

RAVLT measures	Description
<i>Primary measures</i>	
Trial 1	Number of correctly recalled words from list A after first learning trial
Trial 2	Second learning trial
Trial 3	Third learning trial
Trial 4	Fourth learning trial
Trial 5	Fifth learning trial
List B	Free recall of list B
Trial 6	Recall of list A without renewed presentation
Trial 7	Thirty-minute delayed recall of list A
<i>Derived measures</i>	
Trials 1–5 total learning	Σ (Trial 1, Trial 2, Trial 3, Trial 4, Trial 5)
Learning over trials	(Trials 1–5 total—(Trial 1*5))
Learning rate ^a	(Trial 5—Trial 1)
Proactive inhibition ^b	(Trial 1—list B)
Retroactive inhibition ^b	(Trial 5—Trial 6)
Long-term percentage retention	(100 * (Trial 7/Trial 5))

Note: Σ = sum; primary measures are reported in order of administration;

^aPositive score on learning rate indicate that more words were repeated at Trial 5 than Trial 1.

^bPositive score indicate inhibition effect, that is, more words were recalled in Trial 1 compared to List B, or more words recalled in Trial 5 compared to Trial 6.

primary RAVLT measures was suitable for multivariate analysis. Correlations between primary RAVLT measures and demographical variables are presented in [Appendix A.2.1](#).

A preliminary multivariate regression model with predictors age, age², education, education², age*education interaction, sex, sex*age interaction, trial, trial*age interaction, trial*education interaction, trial*sex interaction, and a dummy coded variable accounting for cohort-effects was fitted. Age and education were mean-centered to avoid bias due to multicollinearity and improve interpretation of coefficients. Trial was dummy coded with 7 dummies and trial 1 as the reference category. The preliminary model was subsequently simplified and reduced by hierarchically dropping one covariate at a time in a stepwise manner and comparing log-likelihood ratios of models. The model selection process and associated test statistics are presented in [Appendix A.2](#). Maximum likelihood estimation was used because this allows for classical likelihood ratio testing of nested models (i.e., directly comparing simpler models with complex models). If the simplified model with one reduced covariate did not significantly reduce log-likelihood, then the simplified model was preferred and subsequently used as reference model for further simplification. A nominal alpha-level criterion of $\alpha = .01$ was used. Once the mean structure of the model could not be simplified further without deterioration, the correlation structure of the model was attempted simplified using a homogenous/heterogeneous compound symmetry (CS) and a first-order autoregressive covariance structure (AR (1)). Results indicated that the default unstructured covariance matrix provided the best fit to the data. Once adequate mean structure and covariance structures were obtained, estimates were re-calculated using restricted maximum likelihood (REML), which may reduce small sample bias (Van der Elst et al., 2017; Verbeke & Molenberghs, 2009).

For the derived RAVLT variables, we fitted conventional univariate multiple regression models that were assessed for linear, nonlinear and interaction effects of age, education, and sex. These predictors were included if they significantly improved model fit ($p < .05$). Histograms and QQ-plots of standardized residuals indicated slight deviations from normality for the measure long-term percentage retention (LTPR), and some caution is advised when interpreting extreme scores (e.g., $T < 30$) for this measure. Normative models for the secondary variables are provided in [Table 4](#). We assessed all normative measures for influential cases and outliers that might disturb or unduly influence normative measures. Cases deemed highly influential and abnormal were excluded from analysis to ensure the validity of normative estimates. The variables proactive inhibition and retroactive inhibition were non-normally distributed and had no significant association with demographic variables. We therefore calculated the inverse cumulative distribution based on the performance of the entire normative sample ($n=244$) for these measures. Raw scores and corresponding percentiles are provided in [Table 5](#). All analyses were conducted using the Statistical Package for Social Sciences (SPSS) version 28, JASP version 0.16.1 (JASP Team, 2022), and R version 3.6.2 (R Core Team, 2020).

Calculating normative performance using regression-based norms

Three steps are required for calculating the normative performance: (1) estimating the predicted performance using regression coefficients, (2) subtracting the actual observed score from the predicted score, (3) standardization. Firstly, because age and

years of education was mean centered for all analyses, they must be calculated relative to the age ($M=64.3$), and years of education ($M=12.7$), in the normative sample (Table 1). Every full year of formal education is counted, excluding degrees of the same level. For instance, a participant could reach 24 years of education by 13 years basic schooling, a professional degree of 6 years and a Ph.D. position intended for 5 years. Then, predicted performance is calculated applying the coefficients in Tables 3 and 4. Regression coefficients from the multivariate regression model are applied using the following formula: [Intercept + (individual sex*sex coefficient) + (age centered*age coefficient) + (years of education centered*education coefficient) + (coefficient for Trial n) + (years of education centered * coefficient for education for Trial n) + (individual sex * sex coefficient for Trial n)]. This produces a predicted score based on individual demographics. The predicted score is then subtracted from the individual obtained score. Lastly, the normative score is standardized to the Z-scores following: [Obtained score – predicted scaled score/standard deviation of the residuals obtained from the regression = Z-score]. As customary, Z-scores were further converted to T-scores with a mean of 50 and standard deviation of 10 by [$T=Z * 10 + 50$].

As an example, suppose that a 70-year-old female with 15 years of education remembered 10 words on trial 2. Age centered equals 5.7 [$= 70 - 64.3$] and years of education centered is 2.3 [$15 - 12.7$]. Thus, the predicted score equals 9.1 [$= (5.053 + (1 * 0.761) + (5.7 * -0.041) + (2.3 * 0.128) + 2.457 + (1 * 0.540) + (2.3 * 0.097))$]. The standardized residual for Trial 2 is 1.813. So, the T-score is 55 [$= (((10 - 9.1)/1.813) * 10) + 50$].

Table 3. Coefficients from multivariate regression for normative adjustments on the primary variables from the RAVLT based on 244 healthy adult participants.

Parameter	<i>b</i>	<i>b</i> 95 % CI [LL, UL]	<i>s.e.</i>	<i>t</i>	<i>p</i>	<i>SD residual</i>
Intercept	5.053	[4.750, 5.356]	0.155	32.643	<.001	1.589
Age	-0.041	[-0.065, -0.017]	0.012	-3.318	.001	
Education	0.128	[0.066, 0.189]	0.031	4.076	<.001	
Sex	0.761	[0.357, 1.166]	0.206	3.692	<.001	
Trial 2	2.457	[2.140, 2.773]	0.161	15.218	<.001	1.813
Trial 3	3.826	[3.458, 4.193]	0.187	20.422	<.001	2.068
Trial 4	5.099	[4.694, 5.504]	0.207	24.655	<.001	2.167
Trial 5	5.591	[5.183, 5.999]	0.208	26.876	<.001	2.119
Trial 6	3.282	[2.765, 3.798]	0.264	12.454	<.001	2.677
Trial 7	3.151	[2.643, 3.658]	0.259	12.171	<.001	2.695
List B	0.181	[-0.174, 0.536]	0.181	1.001	.317	1.673
Edu*Trial 2	0.097	[0.033, 0.160]	0.032	2.984	.003	
Edu*Trial 3	0.122	[0.049, 0.196]	0.038	3.253	.001	
Edu*Trial 4	0.120	[0.039, 0.202]	0.042	2.898	.004	
Edu*Trial 5	0.158	[0.076, 0.240]	0.042	3.777	<.001	
Edu*Trial 6	0.212	[0.109, 0.316]	0.053	4.010	<.001	
Edu*Trial 7	0.230	[0.128, 0.332]	0.052	4.425	<.001	
Edu*List B	0.053	[-0.018, 0.124]	0.036	1.458	.145	
Sex*Trial 2	0.540	[0.120, 0.961]	0.215	2.518	.012	
Sex*Trial 3	0.620	[0.132, 1.108]	0.249	2.489	.013	
Sex*Trial 4	0.318	[-0.221, 0.857]	0.275	1.155	.248	
Sex*Trial 5	0.571	[0.029, 1.113]	0.277	2.063	.039	
Sex*Trial 6	0.900	[0.213, 1.587]	0.350	2.569	.010	
Sex*Trial 7	0.712	[0.037, 1.387]	0.344	2.069	.039	
Sex*List B	-0.415	[-0.887, 0.057]	0.241	-1.722	.085	

Notes: Intercept represents reference category Trial 1; *b* = unstandardized beta coefficient; *s.e.* = standard error of the unstandardized beta coefficient; *SD residual* = standard deviation of the residuals; Sex was coded (0 = male, 1 = female); Age and Education were mean centered, thus Age = (calendar age – 64.3); Education/Edu = (the number of years of education obtained – 12.7).

Table 4. Coefficients from multiple regressions for derived RAVLT measures based on 244 healthy adult participants.

Parameter	<i>b</i>	<i>b</i> 95 % CI [LL, UL]	<i>s.e.</i>	<i>t</i>	<i>p</i>	Partial <i>R</i> ²	Adj. <i>R</i> ²	<i>SD</i> <i>residual</i>
Trials 1–5 total intercept	42.269	[40.752, 43.839]	0.783	53.988	<.001		.300	7.982
Trials 1–5 total age	–0.269	[–0.423, –0.116]	0.078	–3.458	<.001	.048		
Trials 1–5 total education	1.095	[0.782, 1.409]	0.159	6.888	<.001	.165		
Trials 1–5 total sex	5.854	[3.795, 7.913]	1.045	5.601	<.001	.116		
LTPR intercept	77.377	[75.289, 79.466]	1.060	72.983	<.001		.073	16.493
LTPR age	–0.427	[–0.741, –0.113]	0.159	–2.682	.008	.029		
LTPR education	1.054	[0.408, 1.700]	0.328	3.215	.001	.041		
LOT intercept	16.972	[15.685, 18.259]	0.653	25.977	<.001		.068	6.699
LOT education	0.479	[0.221, 0.738]	0.131	3.652	<.001	.052		
LOT sex	2.152	[0.440, 3.863]	0.869	2.477	.014	.025		
LR intercept	5.591	[5.181, 6.001]	0.208	26.877	<.001		.064	2.133
LR education	0.158	[0.076, 0.240]	0.042	3.777	<.001	.056		
LR sex	0.571	[0.026, 1.116]	0.277	2.063	.040	.017		

Notes: LTPR, long-term percentage retention ($100 * (\text{Trial 7}/\text{Trial 5})$); LOT, Learning over trials (Trials 1–5 total—(Trial 1*5)); LR=learning rate (Trial 5—Trial 1); *b*=unstandardized beta coefficient; *s.e.* = standard error of the unstandardized beta coefficient; *SD* residual=standard deviation of the residuals; Sex was coded (0=male, 1=female); Age and Education were mean centered, thus Age = (calendar age—64.3); Education = (the number of years of education obtained—12.7).

Comparison of the proposed norms to published norms

T-scores on primary RAVLT measures and trials 1–5 total were calculated for the independent comparison group following the method described in the previous passage. Each participant in the independent comparison group was assigned two different demographically adjusted *T*-scores; one set of *T*-scores using our proposed norms; one set applying norms from Stricker et al. (2021). Multiple regression analyses on *T*-scores were performed to investigate if the predictors sex, age, or education explained variance in *T*-scores. Because *T*-scores should already be adjusted for differences in age, education, and sex a significant result implies that *T*-scores were not adequately corrected for these demographic variables. To reduce error due to chance capitalization, a nominal alpha criterion level of $\alpha = <.01$ for omnibus ANOVAs were used for all analyses. Coefficients related to significant ANOVAs were then interpreted following a conventional α -level criterion of $p <.05$. Plots comparing *T*-scores produced by norms for trial 7 and fitted lines based on predictors age, education and sex are presented in Figure 1.

Norm calculator

The proposed norms are available in a free web-based tool that computes the regression equations. To obtain normative *T*-scores for both RAVLT measures the user simply needs to enter valid demographic values (sex, age, and years of education) and raw-scores from the RAVLT trials. The tool is implemented as a self-contained HTML/Javascript webpage, available at (<https://uit.no/ressurs/uit/cerad/ravlt-calc.html>) and is released as open source at (<https://github.com/DDI-NO/RAVLT-calc>) under Apache License, version 2.0.

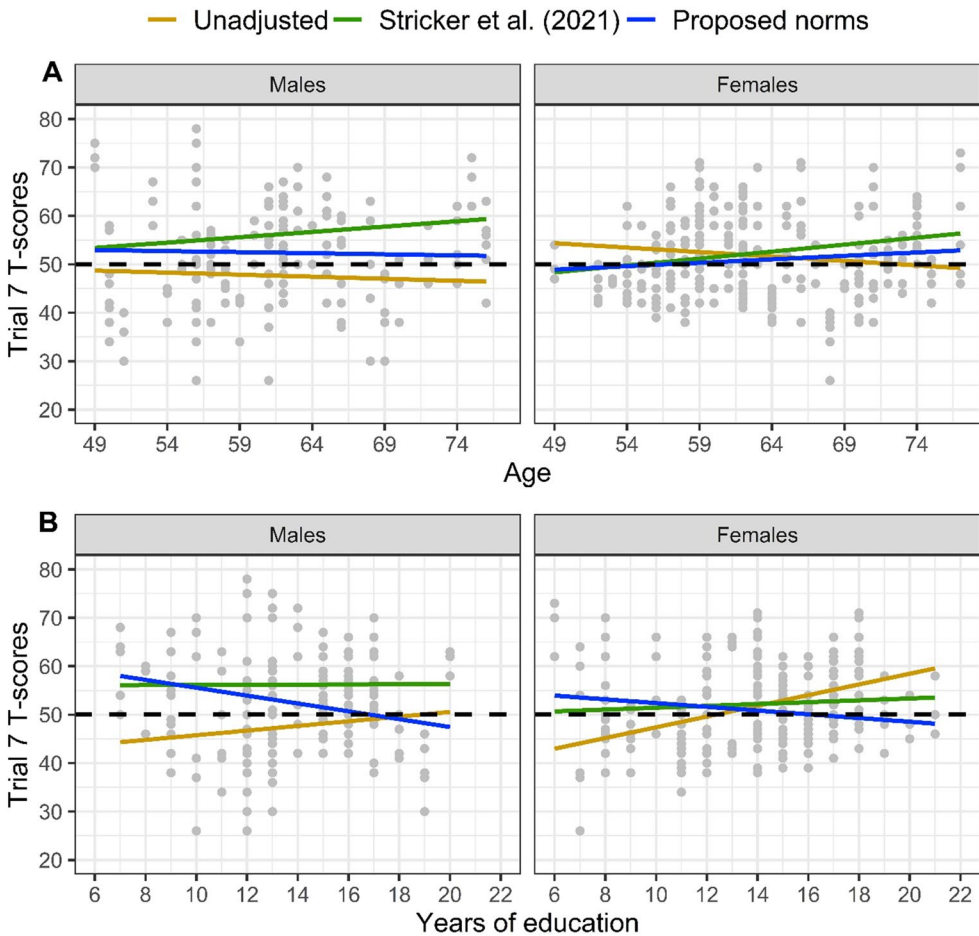


Figure 1. Linear plots of trial 7T-scores computed with Stricker et al. (2021) norms, unadjusted scores and proposed norms.

Test-retest reliability

A sub-set of the normative sample ($n=98$) had available follow-up assessments on the RAVLT allowing for test-retest reliability analysis. The test-retest sample consisted of 65 women (66%) and 33 men (34%) with a mean age of 66.5 years old ($SD=6.6$) and 12.5 ($SD=3.2$) years of education. None of the included participants in the test-retest sample progressed to MCI, dementia or reported symptoms associated with SCD. The average time between assessments was 2.55 years ($SD=0.53$). Intraclass correlation (ICC) estimates and 95% CIs were calculated based on a single rating, absolute-agreement two-way mixed-effects model. Values less than 0.5 indicate poor reliability, 0.5–0.75 moderate reliability and 0.75–0.9 indicate good reliability (Koo & Li, 2016).

Ethics

The Norwegian Regional committees for medical and health research ethics (REK) approved the DDI project from which the current study draws upon. Guidelines in

Helsinki declaration of 1964 (revised 2013) and the Norwegian Health and Research Act were followed. The Gothenburg MCI study was approved by the local ethics committee and conducted in accordance with the Helsinki declaration. All participants gave written informed consents, including the right to withdraw and potential risks and rewards.

Results

Effects of demographics on the RAVLT test performance

The final multivariate model included significant effects of age, education, and sex across all RAVLT trials. As shown in Table 3, higher age was related to lower scores and more years of education were associated with more words recalled on the RAVLT. As expected, participants recalled more words with repeated presentation of the word list as reflected in the coefficients for trials 2–5, where the reference category is trial 1. Results showed that female participants on average recalled 0.76 more words compared to men on trial 1, adjusted for differences in age and education. Furthermore, the effects of education and sex, but not age, differed for subsequent trials of the RAVLT. The interaction term education * trial suggests that the effect of education increased for later parts of the RAVLT. The effect of education was strongest for trial 6 reflecting immediate memory recall ($b=0.21$) and trial 7 which reflected 30 min delayed recall ($b=0.23$). Similarly, the interaction term sex * trial indicate that the difference between men and women was most pronounced on trial 6 ($b=0.90$) and trial 7 ($b=0.71$), where women on average remembered 1.62 and 1.43 more words compared to men. List B, reflecting immediate recall of the novel word list B, did not differ significantly from the reference category trial 1. The effects of education and sex did not differ significantly on list B compared to trial 1.

On the derived measure trials 1–5 total (i.e., sum of words correctly recalled in trials 1–5), there were significant effects of age, sex, and education comparable to the observed effects on trials 1 to 5 separately. On the derived measure long-term percentage retention (LTPR), reflecting the amount of previously learned words on trial 5 retained after a 30-minute delay, lower age ($b=-0.43$), and higher education ($b=1.05$) were significantly related to higher percent retained words (Table 4). We included two measures reflecting learning between trial 1 and trial 5 on the RAVLT, namely learning over trials (LOT) and learning rate (LR). On both measures, higher education predicted increased learning from trial 1 to trial 5 ($b=0.48$; $b=0.16$) and women attained significantly higher scores on learning measures than men ($b=2.15$; $b=0.57$). Lastly, on the measures proactive and retroactive inhibition, we found no significant effect of sex, age, or education. Because these variables followed a non-normal distribution, we report percentiles based on the inverse cumulative distribution of the normative sample (Table 5).

Adjustment of demographics using published norms

Results from multiple regression analysis on demographically adjusted *T*-scores applying norms from Stricker et al. (2021) indicated significant effects of age, education, and sex in the independent comparison group. Omnibus ANOVAs indicated that norms

Table 5. Raw scores to percentile ranks based on the inverse cumulative distribution of the normative sample ($n=244$).

Percentile rank	Retroactive inhibition	Proactive inhibition
2	7	4
5	6	3
10	4–5	2
25	3	1
50	2	0
75	1	–1
90	0	–2
95	–1	–3

Note: Positive scores indicate inhibition effect, i.e., more words were recalled in Trial 1 compared to List B (proactive inhibition), or more words recalled in Trial 5 compared to Trial 6 (retroactive inhibition).

from Stricker et al. (2021) did not adequately adjust for demographics on trial 7 ($F(3, 141) = 5.563, p = .001$), trial 4 ($F(3, 139) = 6.517, p = <.001$), list B ($F(3, 140) = 4.690, p = .004$), and trials 1–5 total ($F(3, 140) = 3.379, p = .006$). As shown in Table 6, adjusted R^2 values indicated that age, sex, and education explained 10.4% of the total variance on trial 4, 8.7% on trial 7, 7.2% on list B, and 6.6% on trials 1–5 total. Stricker et al. (2021) norms did not adequately correct for the effect of age and sex on these trials. On trial 4, list B, trial 7, and trials 1–5 total, female participants were on average estimated 6.9, 4.6, 4.2, and 3.9, T -scores lower than males, respectively, and higher age predicted higher T -scores in all analyses. Omnibus ANOVAs indicated that the current proposed norms adequately adjusted for demographics in the independent comparison group on all measures. However, as shown in Figure 1, there was a tendency for faulty adjustment of education on trial 7, especially for males. In fact, the coefficient for education was significant ($b = -0.53, p = .021$), although omnibus ANOVAs indicate that the combined effect of predictors was not significant.

Test–retest reliability

Trials 6, 7, and trials 1–5 total, showed the best reliability in the follow-up sample, indicating moderate to good reliability. Trials 1 to 5 all had poor and poor-to-moderate reliability. Out of the derived measures, reliability estimates varied from poor to moderate for some trials, with retroactive inhibition and long-term percent retention showing the best reliability.

Discussion

Effects of demographics on the primary measures

We present normative data on the RAVLT based on the performance of a healthy control group from 49 to 79 years from Norway and Sweden ($n=244$). The effect of age in this study stands out as small compared to some previous studies, which all have quoted age as the best predictor for performance (Bezdicsek et al., 2014; Cavaco et al., 2015; Messinis et al., 2016; Stricker et al., 2021). On the combined measure trials 1–5 total, age explained merely 4.8% of the total variance in scores, compared

to 16.5% from education and 11.6% from sex. In other words, in this sample consisting of participants between 49 and 79 years from Norway and Sweden, we found less difference between the younger and elderly participants than expected from other studies. The weak effect of age might be due to the narrower age range comprised solely of middle-aged to elderly adults. Furthermore, the effect of age was the same for different trials on the RAVLT. The effect of age was the same for the initial learning trials as for the 30-minute delayed recall, which is consistent with some (Bezdicek et al., 2014; Messinis et al., 2016; Stricker et al., 2021), but not all studies (Boenniger et al., 2021; Cavaco et al., 2015; Lavoie et al., 2018). Weak effects of age in normative scores are not necessarily a weakness, as some studies have indicated that age-related deterioration might reflect undetected preclinical Alzheimer's disease and other pathological processes (Harrington et al., 2018; Yu et al., 2015).

Women outperformed men on all primary RAVLT trials, and the difference was greatest on trial 6 and trial 7, which has also been demonstrated in the previous studies (Sundermann et al., 2017, 2016). Stricker et al. (2021) argued for the necessity of demographically adjusted *T*-scores that incorporate sex. Their results demonstrated that women significantly outperformed men, and that failure to adjust for sex caused underestimation of amnesic mild cognitive impairment (aMCI) for women and overestimation for men. Previous studies of sex differences on the RAVLT specifically, and verbal memory in general, have found that women outperform men (Asperholm et al., 2019; Van Der Elst et al., 2005) even in samples with Alzheimer's pathology (Sundermann et al., 2017; 2016). As such, our results contribute to the collection of the previous studies that indicate the importance of adjusting for sex on the RAVLT. Despite women performing better than men on trial 7, we found no significant difference on long-term percent retention (LTPR). This suggests that the difference observed between men and women on trial 7 was mainly due to women successfully learning more words on the initial learning trials, and not better retainment of previously learned material per se. Indeed, women were able to learn more words on trial 1, reflecting attentional ability (Woodard, 2006, pp. 105–142), but also amassed more words over the subsequent trials, as reflected in the secondary measures learning over time (LOT) and total learning (TL, Ivnik et al., 1992; Vakil et al., 2010).

Previous studies in samples with comparable educational composition have generally found that education explained a substantial proportion of the variance in scores, but less so than observed in this Scandinavian sample (Bezdicek et al., 2014; Messinis et al., 2016; Stricker et al., 2021; Van Der Elst et al., 2005). This might be due to cultural differences between cohorts, possibly reflecting differences in the educational system and availability of education, or simply variation due to the estimation method. We entered education as a continuous predictor in all analyses and included participants with an extensive range of educational attainment. It is not feasible to provide a conclusive explanation for the difference between norms, particularly in terms of cultural differences, but this highlights the importance of locally sourced norms from a suitable sample that resembles the intended population. Of note, we have previously shown that on the Trail Making Test (TMT), Scandinavians with high education attainment were over-penalized (i.e., received too low *T*-scores) when applying norms from a North American sample by Heaton et al. (2004) and Espenes et al. (2020). On the other hand, Lorentzen et al. (2021) demonstrated that Scandinavians

with high educational attainment received too high *T*-scores compared to the expected normative mean on the controlled oral word association test (COWAT FAS), thus indicating that norms under-adjusted for the effects of education. In sum, we argue that local norms are necessary as the results from the current study suggests that education was more closely related to performance on the RAVLT; the effect of age was smaller; and previous investigations have found foreign norms to inadequately adjust for education when applied in a Scandinavian sample.

Evaluation of norms in an independent comparison group

A key objective of this study was to assess if the proposed norms sufficiently corrected for demographics in an independent comparison group and compare performance with norms from Stricker et al. (2021). Norms from Stricker et al. (2021) adequately corrected for demographics on all RAVLT trials, except trial 4, list B, trial 7 and trials 1–5 total (Table 6). The unstandardized coefficients for age were positive, indicating that increases in age were related to increased *T*-scores. Also, female participants were on average estimated about half a standard deviation lower *T*-scores than males. This suggests that both the generally unfavorable effect of higher age on RAVLT performance, and the difference between men and women, was exaggerated when applied in the independent comparison group. As shown in Table 6, the current proposed norms adequately adjusted for age, education, and sex on all RAVLT trials in the independent comparison group. Regarding education, norms from Stricker et al. (2021) adequately adjusted in all trials. From Figure 1 it is apparent that the current proposed norms produced *T*-scores on trial 7 that exhibited some under-adjustment, especially for males with low levels of education. Although the omnibus ANOVA indicate that the model was insignificant, the individual coefficient for education on trial 7 was significant ($p = .021$; Table 6). This likely stems from sample characteristics; the independent comparison group consisted of very few male participants with lower levels of education that displayed results that exceeded normative expectations. As such, we cannot guarantee the external validity of these results. However, we believe they provide some indication of the norms' ability to adjust in a Scandinavian sample and are valuable for direct comparison of normative adjustments.

Failure to adequately correct for demographics can lead to faulty estimates of the participants' performance, thus influencing the rate of correctly diagnosed patients with amnesic mild cognitive impairment (Stricker et al., 2021). In the normative sample and the independent comparison group females outperformed men on the RAVLT. Applying norms from Stricker et al. (2021) exaggerated the sex difference on the RAVLT such that males had higher *T*-scores than female participants. Over- or underestimation of performance on the RAVLT may result in missed treatment opportunities or unnecessary treatment, which may negatively affect quality of life (Stricker et al., 2021). Failure to adjust for age and education is most apparent in the end ranges of predictors, that is, for the youngest and oldest and individuals with either very low or very high levels of education. For example, a 68-year-old male with 19 years of formal education enrolled in the independent comparison group remembered 5 words on the 30-min delay on the RAVLT (trial 7). Applying norms from



Table 6. Results from multiple regression analysis on T-scores with predictors age, years of education, and sex in the independent comparison group ($n = 145$).

Variable	Predictor	Stricker et al.'s (2021) norms				Scandinavian norms			
		b	p	Partial R^2	Adj. R^2	b	p	Partial R^2	Adj. R^2
Trial 1	Intercept	53.573	<.001	.020	.043	51.989	<.001	.015	.015
	Age	0.247	.091	.038		0.205	.151	.013	
	Education	0.716	.020	.014		0.402	.181	.013	
	Sex	-2.815	.164			-2.704	.173		
Trial 2	Intercept	55.343	<.001	.018	.022	52.417	<.001	<.001	.011
	Age	0.229	.115	.005		0.041	.776	<.001	
	Education	0.263	.386	.024		-0.195	.523	.003	
	Sex	-3.740	.065		.015	-4.072	.046	.028	
Trial 3	Intercept	54.799	<.001	.025		52.576	<.001	.001	-.013
	Age	0.276	.060	.007		0.064	.650	<.001	
	Education	0.302	.332	.008		-0.105	.721	<.001	
	Sex	-2.177	.283		.104*	-1.758	.369	.006	
Trial 4	Intercept	58.029	<.001	.031		53.888	<.001	<.001	-.020
	Age	0.262	.037	.006		0.039	.879	<.001	
	Education	0.229	.383	.10		-0.177	.741	<.001	
	Sex	-6.852	<.001		.013	-0.922	.796	<.001	
Trial 5	Intercept	53.499	<.001	.027		50.150	<.001	<.001	.028
	Age	0.242	.049	<.001		0.065	.564	<.001	
	Education	-0.024	.925			-0.606	.012	.045	
	Sex	-1.673	.324	.007		-0.018	.991	<.001	
Trial 6	Intercept	54.875	<.001	.017	.025	52.456	<.001	.002	.006
	Age	0.158	.127	.009		-0.064	.572	.014	
	Education	0.246	.255	.026		-0.337	.157	.010	
	Sex	-2.731	.057			-1.894	.229		
Trial 7	Intercept	56.363	<.001	.051	.087*	52.396	<.001	.001	.027
	Age	0.267	.007	.005		0.048	.656	.037	
	Education	0.179	.384	.066		-0.527	.021	.008	
	Sex	-4.215	.002			-1.566	.296		
List B	Intercept	57.575	.001	.062	.072*	50.539	<.001	.036	.022
	Age	0.432	.003	.005		0.327	.025	.001	
	Education	0.253	.395	.037		-0.047	.879	<.001	
	Sex	-4.555	.022			-2.238	.262	.009	
Trials 1–5 Total	Intercept	55.839	<.001	.055	.066*	52.227	<.001	.022	.015
	Age	0.377	.005	.012		0.229	.078	.001	
	Education	0.353	.205	.031		-0.104	.703	.001	
	Sex	-3.885	.035			-2.505	.164	.014	

Notes: b = unstandardized regression coefficient; p = p-value; partial R^2 = explained variance of predictor variable; Adj. R^2 = explained variance of combined predictor variables
*Omnibus ANOVA was significant ($p < .01$), and we report F and P-values for significant models in text; age and education was mean centered.

Stricker et al. (2021) the calculated T -score is $T=43$ compared to $T=30$ applying the current proposed norms. Thus, applying diagnostic criteria for amnesic mild cognitive impairment (aMCI, Albert et al., 2011; Bondi et al., 2014), this could have implications for correctly diagnosing aMCI and providing adequate treatment.

Test–retest reliability

Test–retest reliability is important for neuropsychological tests that are used to inform decisions on the cognitive status of patients at the time of assessments and their likely functioning in the future (Sherman et al., 2011). Trial 1, LOT and LR showed poor reliability (Table 7). This might be expected on attentional measures that typically show lower test–retest reliability as attention is considered a “changeable trait” (Sherman et al., 2011) compared to verbal memory, which may be regarded “trait-like” and stable in healthy participants. Thus, clinicians should exercise caution interpreting these measures in isolation. Instead, clinicians concerned with the reliability of test scores are recommended to use trials 1–5 total as a measure of acquisition, attention, and learning which showed moderate to good reliability. Both trial 6 and trial 7 showed moderate-to-good reliability, and LTPR showed poor-to-moderate reliability. The same pattern of test–retest reliability was reported by Stricker et al. (2021), though our reliability estimates were slightly lower overall. This is likely due to the longer test–retest interval in this study (2.5 years compared to 1.5 years) and small sample size ($n=98$) for the follow-up group, thus inflating the associated 95% confidence intervals.

Effects of demographics on the derived measures

We provided norms for retroactive and proactive inhibition measures, which might have utility in specific clinical samples burdened with executive deficits. Proactive inhibition refers to the reduced ability to learn new material due to interference from

Table 7. Test–retest reliability of RAVLT measures based on a subset of the normative sample ($n=98$).

Measure	ICC	95% CI [LL, UL]
Trial 1	0.324	[0.135, 0.491]
Trial 2	0.504	[0.335, 0.642]
Trial 3	0.511	[0.343, 0.647]
Trial 4	0.457	[0.279, 0.604]
Trial 5	0.560	[0.407, 0.682]
List B	0.549	[0.394, 0.674]
Trial 6	0.749	[0.646, 0.825]
Trial 7	0.712	[0.598, 0.797]
Trials 1–5 Total	0.659	[0.528, 0.759]
Learning over trials	0.174	[0.028, 0.363]
Learning rate	0.178	[−0.021, 0.364]
Proactive inhibition	−0.030	[−0.228, 0.17]
Retroactive inhibition	0.378	[0.193, 0.537]
Long-term percent retention	0.532	[0.372, 0.661]

Note: ICC, intraclass correlation coefficient.

the previously learned material, and is derived on the RAVLT by comparing performance on list B to trial 1. On the other hand, retroactive inhibition refers to the reduced ability to recall the previously learned material after inference has occurred (list B) and is measured by comparing performance on trial 5 with trial 6. We did not observe a significant difference between list B and trial 1 (Table 3), thus indicating no significant proactive inhibition in the normative sample on average. In line with the previous normative studies on cognitively healthy adults, we found no significant relationship between proactive inhibition and age (Boenniger et al., 2021; Vakil et al., 2010) or sex (Boenniger et al., 2021), or education. Proactive inhibition has been shown to be deficient in patients with frontal lobe lesions on a paired association test compared to healthy controls (Depue, 2012; Shimamura et al., 1995). Reduced performance on inhibition tasks has been associated with deficits in inhibition, response competition, deficits in source memory, and over-activation of irrelevant memory items (Vakil et al., 2010). As such, it may be expected to find significant deficits in inhibition (either proactive or retroactive) in clinical samples. However, as far as we are aware, there have been no studies comparing performance on these measures in samples with MCI or AD dementia on the RAVLT. Some retroactive inhibition appears to be normal, as participants on average remembered about 2 fewer words on trial 6 than trial 5 (Tables 1 and 3). Patients with schizophrenia have been shown to be susceptible to retroactive inhibition, owing to executive demands associated with retroactive inhibition. Specifically, the ability to; inhibit responses, verbal fluency to govern retrieval of target items; and memory of temporal order (Torres et al., 2001). Boenniger et al. (2021) and Vakil et al. (2010) found a small effect of age on retroactive inhibition, and Boenniger reported that men presented slightly more retroactive inhibition than women. Nevertheless, we found no significant effect of age, sex, or education on retroactive inhibition. Due to the lack of association to demographic variables, we simply report percentile ranks on these measures for clinicians to inform decisions on abnormal/normal performance.

Limitations

Some limitations of the current study are to be addressed. Firstly, participants were not formally screened for auditory deficits which might influence performance on the RAVLT. However, all participants with hearing aids were instructed to use these when applicable. The normative sample from which norms were computed was not a randomized sample of the Norwegian and Swedish population. We therefore cannot guarantee that this sample reflects the population in general. However, this limitation is not specific to this study. Still, it remains a common issue in the normative literature, with exceptions such as the Mayo normative study (Stricker et al., 2021) and the Rhineland study (Boenniger et al., 2021). Also, compared to some previous studies, the normative sample of the current study is relatively small, which influences the degree of certainty that a normative score reflects the true population parameters, especially for extreme scores (e.g., 1.5 *SD* below the sample mean, Crawford & Garthwaite, 2008; Oosterhuis et al., 2016). Lastly, while the experience of SCD is generally considered a normal and benign condition in an aging population (Bassett &

Folstein, 1993; Hessen et al., 2017), it is nevertheless a known risk-factor of neurodegenerative disease (Jessen et al., 2014). However, all included participants were cognitively healthy at the time of analysis, also supported by mean RAVLT scores being largely equivalent to our Normative sample (Table 1).

Conclusion

We propose regression-based test norms for the RAVLT based on a sample of healthy Swedish and Norwegian participants between 49 and 79 years old. A free online norm calculator is offered to improve availability of norms in clinical settings. Test-retest reliability analyses indicated that basic RAVLT trials showed poor-to-moderate reliability, while measures of total learning and verbal memory showed moderate-to-good reliability. Our results indicate that the current proposed norms successfully adjust for age, education, and sex in the independent comparison group. Norms from Stricker et al. (2021) overestimated the effect of age and difference between sexes on parts of the RAVLT. Notably, the failure to adequately adjust for demographical variables on the 30-min delayed recall (trial 7) might have implications for correctly diagnosing amnesic mild cognitive impairment (aMCI) in Scandinavian adults and elderly.

Acknowledgments

We thank Ragna Espenes and Ramune Grambaite for clinical examinations and help with the project.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the University of Tromsø - the Arctic University of Norway; the Norwegian Research Council, JPND/PMI-AD under grant number 311993; and Helse Nord under grant number HNF1540-20. Additional support was received from the Sahlgrenska University Hospital, the Swedish Research Council, Swedish Brain Power, the Swedish Dementia Foundation, the Swedish Alzheimer's Foundation, Stiftelsen Psykiatriska forskningsfonden, and Konung Gustaf V:s och Drottning Victorias Frimurarestiftelse, Demensförbundet, Helse Nord RHF, and EU Joint Programme - Neurodegenerative Disease Research. The funding sources were not involved in the drafting of this manuscript.

ORCID

Jacob Espenes  <http://orcid.org/0000-0002-2383-5348>
Ingvild Vøllo Eliassen  <http://orcid.org/0000-0003-1288-6032>
Fredrik Öhman  <http://orcid.org/0000-0002-6872-4481>
Knut Waterloo  <http://orcid.org/0000-0003-3447-8312>
Ingrid Myrvoll Lorentzen  <http://orcid.org/0000-0002-9942-7594>
Tormod Fladby  <http://orcid.org/0000-0002-9984-9797>
Bjørn-Eivind Kirsebom  <http://orcid.org/0000-0002-1413-9578>

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy of the research participants.

References

- Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B., & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3), 270–279. <https://doi.org/10.1016/j.jalz.2011.03.008>
- Asperholm, M., Nagar, S., Dekhtyar, S., & Herlitz, A. (2019). The magnitude of sex differences in verbal episodic memory increases with social progress: Data from 54 countries across 40 years. *PLoS One*, 14(4), e0214945. <https://doi.org/10.1371/journal.pone.0214945>
- Bassett, S. S., & Folstein, M. F. (1993). Memory complaint, memory performance, and psychiatric diagnosis: A community study. *Journal of Geriatric Psychiatry and Neurology*, 6(2), 105–111. <https://doi.org/10.1177/089198879300600207>
- Belleville, S., Fouquet, C., Hudon, C., Zomahoun, H. T. V., & Croteau, J. Consortium for the Early Identification of Alzheimer's, d.-Q. (2017). Neuropsychological measures that predict progression from mild cognitive impairment to Alzheimer's type dementia in older adults: A systematic review and meta-analysis. *Neuropsychology Review*, 27(4), 328–353. Retrieved from <https://doi.org/10.1007/s11065-017-9361-5>
- Bezdicek, O., Stepankova, H., Moták, L., Axelrod, B. N., Woodard, J. L., Preiss, M., Nikolai, T., Růžička, E., & Poreh, A. (2014). Czech version of Rey Auditory Verbal Learning Test: Normative data. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, 21(6), 693–721. <https://doi.org/10.1080/13825585.2013.865699>
- Boake, C. (2000). Edouard Claparede and the auditory verbal learning test. *Journal of Clinical and Experimental Neuropsychology*, 22(2), 286–292. [https://doi.org/10.1076/1380-3395\(200004\)22:2;1-1;FT286](https://doi.org/10.1076/1380-3395(200004)22:2;1-1;FT286)
- Boenniger, M. M., Staerk, C., Coors, A., Huijbers, W., Ettinger, U., & Breteler, M. M. B. (2021). Ten German versions of Rey's auditory verbal learning test: Age and sex effects in 4,000 adults of the Rhineland Study. *Journal of Clinical and Experimental Neuropsychology*, 43(6), 637–653. <https://doi.org/10.1080/13803395.2021.1984398>
- Bondi, M. W., Edmonds, E. C., Jak, A. J., Clark, L. R., Delano-Wood, L., McDonald, C. R., Nation, D. A., Libon, D. J., Au, R., Galasko, D., & Salmon, D. P. (2014). Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. *Journal of Alzheimer's Disease: JAD*, 42(1), 275–289. <https://doi.org/10.3233/JAD-140276>
- Cavaco, S., Gonçalves, A., Pinto, C., Almeida, E., Gomes, F., Moreira, I., Fernandes, J. & Teixeira-Pinto, A. (2015). Auditory Verbal Learning Test in a Large Nonclinical Portuguese Population. *Applied Neuropsychology. Adult*, 22(5), 321–331. <https://doi.org/10.1080/23279095.2014.927767> 25580839
- Correia, A. F., & Osorio, I. C. (2014). The Rey Auditory Verbal Learning Test: normative data developed for the Venezuelan population. *Arch Clin Neuropsychol*, 29(2), 206–215. <https://doi.org/10.1093/arclin/act070>
- Crawford, J. R., & Garthwaite, P. H. (2008). On the “optimal” size for normative samples in neuropsychology: Capturing the uncertainty when normative data are used to quantify the standing of a neuropsychological test score. *Child Neuropsychology: a Journal on Normal and Abnormal Development in Childhood and Adolescence*, 14(2), 99–117. <https://doi.org/10.1080/09297040801894709>
- Depue, B. E. (2012). A neuroanatomical model of prefrontal inhibitory modulation of memory retrieval. *Neuroscience and Biobehavioral Reviews*, 36(5), 1382–1399.
- Eckerström, C., Olsson, E., Bjerke, M., Malmgren, H., Edman, Å., Wallin, A., & Nordlund, A. (2013). A combination of neuropsychological, neuroimaging, and cerebrospinal fluid markers predicts

- conversion from mild cognitive impairment to dementia. *Journal of Alzheimer's Disease: JAD*, 36(3), 421–431. <https://doi.org/10.3233/JAD-122440>
- Eliassen, I. V., Fladby, T., Kirsebom, B.-E., Waterloo, K., Eckerström, M., Wallin, A., Bråthen, G., Aarsland, D., & Hessen, E. (2020). Predictive and diagnostic utility of brief neuropsychological assessment in detecting Alzheimer's pathology and progression to dementia. *Neuropsychology*, 34(8), 851–861. <https://doi.org/10.1037/neu0000698>
- Espenes, J., Hessen, E., Eliassen, I. V., Waterloo, K., Eckerström, M., Sando, S. B., Timón, S., Wallin, A., Fladby, T., & Kirsebom, B.-E. (2020). Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years. *The Clinical Neuropsychologist*, 34(sup1), 110–126. <https://doi.org/10.1080/13854046.2020.1829068>
- Estévez-González, A., Kulisevsky, J., Boltes, A., Otermin, P., & García-Sánchez, C. (2003). Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: Comparison with mild cognitive impairment and normal aging. *International Journal of Geriatric Psychiatry*, 18(11), 1021–1028.
- Ewers, M., Walsh, C., Trojanowski, J. Q., Shaw, L. M., Petersen, R. C., Jack, C. R., Feldman, H. H., Bokde, A. L. W., Alexander, G. E., Scheltens, P., Vellas, B., Dubois, B., Weiner, M., & Hampel, H. North American Alzheimer's Disease Neuroimaging Initiative (ADNI). (2012). Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiology of Aging*, 33(7), 1203–1214. e1202. <https://doi.org/10.1016/j.neurobiolaging.2010.10.019>
- Fillenbaum, G. G., van Belle, G., Morris, J. C., Mohs, R. C., Mirra, S. S., Davis, P. C., Tariot, P. N., Silverman, J. M., Clark, C. M., Welsh-Bohmer, K. A., & Heyman, A. (2008). Consortium to establish a registry for Alzheimer's disease (CERAD): The first twenty years. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 4(2), 96–109. <https://doi.org/10.1016/j.jalz.2007.08.005>
- Fladby, T., Pålhaugen, L., Selnes, P., Waterloo, K., Bråthen, G., Hessen, E., Almdahl, I. S., Arntzen, K.-A., Auning, E., Eliassen, C. F., Espenes, R., Grambaite, R., Grøntvedt, G. R., Johansen, K. K., Johnsen, S. H., Kalheim, L. F., Kirsebom, B.-E., Müller, K. I., Nakling, A. E., ... Aarsland, D. (2017). Detecting at-risk Alzheimer's disease cases. *Journal of Alzheimer's Disease: JAD*, 60(1), 97–105. <https://doi.org/10.3233/JAD-170231>
- Harrington, K. D., Schembri, A., Lim, Y. Y., Dang, C., Ames, D., Hassenstab, J., Laws, S. M., Rainey-Smith, S., Robertson, J., Rowe, C. C., Sohrabi, H. R., Salvado, O., Weinborn, M., Villemagne, V. L., Masters, C. L., & Maruff, P. AIBL Research Group. (2018). Estimates of age-related memory decline are inflated by unrecognized Alzheimer's disease. *Neurobiology of Aging*, 70, 170–179.
- Heaton, R., Miller, S., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults*. Psychological Assessment Resources.
- Hessen, E., Eckerström, M., Nordlund, A., Selseth Almdahl, I., Stålhammar, J., Bjerke, M., Eckerström, C., Göthlin, M., Fladby, T., Reinvang, I., & Wallin, A. (2017). Subjective cognitive impairment is a predominantly benign condition in memory clinic patients followed for 6 years: The Gothenburg-Oslo MCI study. *Dementia and Geriatric Cognitive Disorders Extra*, 7(1), 1–14. Retrieved from <https://doi.org/10.1159/000454676>
- Hessen, E., Reinvang, I., Eliassen, C. F., Nordlund, A., Gjerstad, L., Fladby, T., & Wallin, A. (2014). The combination of dysexecutive and amnesic deficits strongly predicts conversion to dementia in young mild cognitive impairment patients: A report from the Gothenburg-Oslo MCI study. *Dementia and Geriatric Cognitive Disorders Extra*, 4(1), 76–85. <https://doi.org/10.1159/000360282>
- Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992). Mayo's older Americans normative studies: Updated AVLT norms for ages 56 to 97. *Clinical Neuropsychologist*, 6(sup001), 83–104. <https://doi.org/10.1080/13854049208401880>
- Ivnik, R. J., Malec, J. F., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1990). The auditory-verbal learning test (AVLT): norms for ages 55 years and older. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 2(3), 304–312. <https://doi.org/10.1037/1040-3590.2.3.304>

- JASP Team (2022). JASP (Version 0.16.3)[Computer software].
- Jessen, F., Amariglio, R. E., Van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., ... Van Der Flier, W. M. (2014). A Conceptual Framework for Research on Subjective Cognitive Decline in Preclinical Alzheimer's Disease. *Alzheimer's & Dementia*, 10(6), 844–852.
- Kirsebom, B.-E., Espenes, R., Hessen, E., Waterloo, K., Johnsen, S. H., Gundersen, E., Botne Sando, S., Rolfseng Grøntvedt, G., Timón, S., & Fladby, T. (2019). Demographically adjusted CERAD wordlist test norms in a Norwegian sample from 40 to 80 years. *The Clinical Neuropsychologist*, 33(sup1), 27–39. <https://doi.org/10.1080/13854046.2019.1574902>.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Lavoie, M., Bherer, L., Joubert, S., Gagnon, J.-F., Blanchet, S., Rouleau, I., Macoir, J., & Hudon, C. (2018). Normative data for the Rey Auditory Verbal Learning Test in the older French-Quebec population. *The Clinical Neuropsychologist*, 32(sup1), 15–28. <https://doi.org/10.1080/13854046.2018.1429670>
- Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment*. (5th ed.). Oxford University Press.
- Lorentzen, I. M., Espenes, J., Hessen, E., Waterloo, K., Bråthen, G., Timón, S., ... Kirsebom, B.-E. (2021). Regression-based norms for the FAS phonemic fluency test for ages 40–84 based on a Norwegian sample. *Applied Neuropsychology: Adult*, 1–10.
- Loring, D. W., Strauss, E., Hermann, B. P., Barr, W. B., Perrine, K., Trenerry, M. R., Chelune, G., Westerveld, M., Lee, G. P., Meador, K. J., & Bowden, S. C. (2008). Differential neuropsychological test sensitivity to left temporal lobe epilepsy. *Journal of the International Neuropsychological Society*, 14(03), 394–400. <https://doi.org/10.1017/S1355617708080582>
- Marqués, N. O., Caro, I. A., Uterga Valiente, J. M., & Rodríguez, S. M. (2013). Normative data for a Spanish version of the Rey Auditory-Verbal Learning Test in older people. *The Spanish Journal of Psychology*, 16, E60. <https://www.cambridge.org/core/journals/spanish-journal-of-psychology/article/abs/normative-data-for-a-spanish-version-of-the-rey-auditoryverbal-learning-test-in-older-people/6E3C9F4A3D9677B29C2AAE630E2DF1CF>.
- Messinis, L., Nasios, G., Mougias, A., Politis, A., Zampakis, P., Tsiamakia, E., Malefaki, S., Gourzis, P., & Papatathanasopoulos, P. (2016). Age and education adjusted normative data and discriminative validity for Rey's Auditory Verbal Learning Test in the elderly Greek population. *Journal of Clinical and Experimental Neuropsychology*, 38(1), 23–39. <https://doi.org/10.1080/13803395.2015.1085496>
- Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment*. : Oxford University Press.
- Molinuevo, J. L., Rabin, L. A., Amariglio, R., Buckley, R., Dubois, B., Ellis, K. A., Ewers, M., Hampel, H., Klöppel, S., Rami, L., Reisberg, B., Saykin, A. J., Sikkes, S., Smart, C. M., Snitz, B. E., Sperling, R., van der Flier, W. M., Wagner, M., & Jessen, F. Subjective Cognitive Decline Initiative (SCD-I) Working Group. (2017). Implementation of subjective cognitive decline criteria in research studies. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, 13(3), 296–311. <https://doi.org/10.1016/j.jalz.2016.09.012>
- Oosterhuis, H. E. M., van der Ark, L. A., & Sijsma, K. (2016). Sample size requirements for traditional and regression-based norms. *Assessment*, 23(2), 191–202. <https://doi.org/10.1177/1073191115580638>
- Powell, J. B., Cripe, L. I., & Dodrill, C. B. (1991). R: A language and environment for statistical computing. *Archives of Clinical Neuropsychology*, 6(4), 241–249. <https://doi.org/10.1093/arc-clin/6.4.241>
- R Core Team (2020). Assessment of brain impairment with the Rey Auditory Verbal Learning Test: A comparison with other neuropsychological measures. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation*. (Vol. 4): Reitan Neuropsychology.
- Rey, A. (1958). *L'examen clinique en psychologie*. [The clinical examination in psychology]. Presses Universitaires De France.

- Ricci, M., Graef, S., Blundo, C., & Miller, L. A. (2012). Using the Rey Auditory Verbal Learning Test (RAVLT) to differentiate Alzheimer's dementia and behavioural variant fronto-temporal dementia. *The Clinical Neuropsychologist*, 26(6), 926–941.
- Sherman, E., Brooks, B. L., Iverson, G. L., Slick, D. J., & Strauss, E. (2011). Reliability and validity in neuropsychology. In *The little black book of neuropsychology* (pp. 873–892): Springer.
- Shimamura, A. P., Jurica, P. J., Mangels, J. A., Gershberg, F. B., & Knight, R. T. (1995). Susceptibility to memory interference effects following frontal lobe damage: Findings from tests of paired-associate learning. *Journal of Cognitive Neuroscience*, 7(2), 144–152.
- Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. American Chemical Society.
- Stricker, N. H., Christianson, T. J., Lundt, E. S., Alden, E. C., Machulda, M. M., Fields, J. A., Kremers, W. K., Jack, C. R., Knopman, D. S., Mielke, M. M., & Petersen, R. C. (2021). Mayo normative studies: Regression-based normative data for the auditory verbal learning test for ages 30-91 years and the importance of adjusting for sex. *Journal of the International Neuropsychological Society*, 27(3), 211–226. <https://doi.org/10.1017/S1355617720000752>
- Sundermann, E. E., Biegon, A., Rubin, L. H., Lipton, R. B., Landau, S., Maki, P. M., for the Alzheimer's, D., & Neuroimaging, I, Alzheimer's Disease Neuroimaging Initiative (2017). Does the female advantage in verbal memory contribute to underestimating Alzheimer's disease pathology in women versus men? *Journal of Alzheimer's Disease: JAD*, 56(3), 947–957. <https://doi.org/10.3233/JAD-160716>
- Sundermann, E. E., Biegon, A., Rubin, L. H., Lipton, R. B., Mowrey, W., Landau, S., Initiative, F., ... t, A, s. D. N (2016). Better verbal memory in women than men in MCI despite similar levels of hippocampal atrophy. *Neurology*, 86(15), 1368–1376. Retrieved from <https://doi.org/10.1212/wnl.0000000000002570>.
- Torres, I. J., Flashman, L. A., O'Leary, D. S., & Andreasen, N. C. (2001). Effects of retroactive and proactive interference on word list recall in schizophrenia. *Journal of the International Neuropsychological Society : JINS*, 7(4), 481–490. <https://doi.org/10.1017/S1355617701744049>
- Vakil, E., Greenstein, Y., & Blachstein, H. (2010). Normative Data for Composite Scores for Children and Adults Derived from the Rey Auditory Verbal Learning Test. *The Clinical Neuropsychologist*, 24(4), 662–677. <https://doi.org/10.1080/13854040903493522>
- Van der Elst, W., Molenberghs, G., van Tetering, M., & Jolles, J. (2017). Establishing normative data for multi-trial memory tests: the multivariate regression-based approach. *The Clinical Neuropsychologist*, 31(6/7), 1173–1187.
- Van Der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., & 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 : JINS*, 11(3), 290–302. <https://doi.org/10.1017/S1355617705050344>
- Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media.
- Wallin, A., Nordlund, A., Jonsson, M., Lind, K., Edman, Å., Göthlin, M., Stålhammar, J., Eckerström, M., Kern, S., Börjesson-Hanson, A., Carlsson, M., Olsson, E., Zetterberg, H., Blennow, K., Svensson, J., Öhrfelt, A., Bjerke, M., Rolstad, S., & Eckerström, C. (2016). The Gothenburg MCI study: Design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow and Metabolism : official Journal of the International Society of Cerebral Blood Flow and Metabolism*, 36(1), 114–131. <https://doi.org/10.1038/jcbfm.2015.147>
- Warrington, E., & James, M. (1991). *VOSP visual object and space perception test battery*. TVTC Thames Valley Test Company.
- Woodard, J. L. (2006). Memory performance indexes for the Rey Auditory Verbal Learning Test. In *The quantified process approach to neuropsychological assessment*. (pp. 105–141). Taylor & Francis.
- Yu, L., Boyle, P. A., Segawa, E., Leurgans, S., Schneider, J. A., Wilson, R. S., & Bennett, D. A. (2015). Residual decline in cognition after adjustment for common neuropathologic conditions. *Neuropsychology*, 29(3), 335–343.

Appendix A

A.1. RAVLT test version and administration procedures in Norwegian and Swedish

Table A.1. Wordlists for RAVLT.

Norwegian		Swedish		English	
List A	List B	List A	List B	List A	List B
Tromme	Bord	Trumma	Skrivebord	Drum	Desk
Gardin	Jeger	Gardin	Polis	Curtain	Ranger
Måne	Fugl	Måne	Fågel	Bell	Bird
Kaffe	Sko	Kaffe	Sko	Coffee	Shoe
Skole	Ovn	Skola	Spis	School	Stove
Foreldre	Fjell	Bror	Berg	Parent	Mountain
Klokke	Briller	Klocka	Glas	Moon	Glasses
Hage	Håndkle	Trädgård	Penna	Garden	Towel
Hatt	Sky	Hatt	Moln	Hat	Cloud
Bonde	Båt	Bonde	Båt	Farmer	Boat
Nese	Lam	Nos	Lamm	Nose	Lamb
Kalkun	Pistol	Kalkon	Pistol	Turkey	Gun
Farge	Blyant	Färg	Handduk	Color	Pencil
Hus	Kirke	Hus	Kyrka	House	Church
Elv	Fisk	Flod	Fisk	River	Fish

Table A.1.2. Norwegian items for RAVLT recognition.

Klokke	Hjem	Håndkle	Båt	Briller
J N	J N	J N	J N	J N
Vindu	Fisk	Gardin	Varm	Strømpe
J N	J N	J N	J N	J N
Hatt	Måne	Blomst	Foreldre	Sko
J N	J N	J N	J N	J N
Låve	Tre	Farge	Vann	Laerer
J N	J N	J N	J N	J N
Jeger	Ballong	Bord	Bonde	Ovn
J N	J N	J N	J N	J N
Nese	Fugl	Gevaer	Rose	Rede
J N	J N	J N	J N	J N
Vaer	Fjell	Fargestift	Sky	Barn
J N	J N	J N	J N	J N
Skole	Kaffe	Kirke	Hus	Tromme
J N	J N	J N	J N	J N
Hånd	Mus	Kalkun	Fremmed	Karamell
J N	J N	J N	J N	J N
Blyant	Elv	Kilde	Hage	Lam
J N	J N	J N	J N	J N

Note. We do not provide normative data for the recognition trial. Correct items from list A are highlighted in bold text.

A.2. Swedish Administration procedures

ADMINISTRERING A1 - A6

En ordlista bestående av 15 ord läses upp för patienten.

Jag kommer att läsa upp en lista med ord för dig, och jag vill att du försöker lägga orden på minnet. När jag har läst listan klart, så vill jag att du säger de ord du kan minnas. Det är många ord, så du kommer inte att kunna minnas alla, men försök minnas så många du kan.

Läs listan i ett tempo av ungefärligen ett ord per sekund. När patienten återger, notera i protokollet vilken ordning orden återges i, samt eventuella upprepningar och konfabulationer. När patienten har varit tyst en stund, fråga om hon minns något mer. En del patienter ger upp snabbt vid denna uppgift och kan behöva uppmuntras att försöka tänka en liten stund till.

Nu kommer jag att läsa den här listan några gånger. Efter varje gång vill jag att du räknar upp de ord du minns. Du ska också ta med de ord som du har sagt tidigare.

Efter den femte retentionen läses en distraktionslista bestående av nya ord upp, och patienten ska återge ord från den nya listan.

Nu kommer jag att läsa en lista med helt nya ord. Även nu vill jag att du försöker att minnas dem, och sedan säga de ord du kan komma ihåg när jag har läst listan färdigt. Den här listan kommer jag bara att läsa en gång.

Patienten uppmanas därefter att återge vad hon nu minns från den första listan. Efter moment A6 går testledaren vidare i protokollet med övriga uppgifter i ca 30 minuter.

ADMINISTRERING A7

Efter ca 30 minuter ombeds patienten igen att dra sig till minnes den första listan.

A.3. Norwegian administration procedures

Administrering liste A, første presentasjon (trial 1).

Jeg vil nå lese opp en liste med ord. Hør nøye etter, for når jeg er ferdig vil jeg at du skal gjenta så mange som du kan huske. Rekkefølgen du sier det i har ingenting å si. Bare prøv å husk så mange du kan.

Liste A, andre presentasjon (trial 2).

Nå vil jeg lese den samme listen med ord igjen og på samme måte vil jeg at du skal gjenta så mange ord som du kan huske, inkludert de ordene du sa første gangen. Rekkefølgen som du sier ordene har ingenting å si, bare gjenta så mange ord du klarer uansett om du sa det første gang.

Gjenta instruksjonen ved behov for trial 3-5.

Direkte etter femte presentasjon skal liste B administreres. Si:

Jeg vil nå lese opp en ny liste med ord, og på samme måte som før skal du prøve huske så mange ord som mulig fra denne nye listen. Rekkefølgen du sier det i har ingenting å si.

Uten fornyet presentasjon av liste A skal testdeltager gjentake liste A på nytt. Si:

Kan du på nytt si alle ordene du husker fra den første listen?

Etter 30 minutter skal testdeltager gjenta liste A for siste gang. Si:

For litt siden leste jeg opp en liste med ord til deg flere ganger og du skulle forsøke lære disse ordene. Kan du gjenta disse ordene en gang til?

A.4. Regression norming procedure

Table A.2. Model selection procedure for the multivariate normative model (n = 244).

Model	Model structure	Cov. Structure	df	-2 log lik difference (G2)	p-value	Ref. model	Qualitative conclusion	BIC	AIC
1	Full	UN	74	-3550.00				7660.66	7247.99
2	Exclude Cohort	UN	72	1.71	.425	1	Exclude Cohort	7647.22	7245.70
3	Exclude edu ²	UN	71	0.004	.951	2	Exclude edu ²	7639.65	7243.71
4	Exclude age ²	UN	70	0.375	.549	3	Exclude age ²	7632.45	7242.08
5	Exclude age*edu	UN	69	0.18	.668	4	Exclude age*edu	7625.05	7240.27
6	Exclude age*sex	UN	68	0.19	.663	5	Exclude age*sex	7617.67	7238.46
7	Exclude trial*age	UN	61	8.23	.313	6	Exclude trial*age	7572.86	7232.69
8	Exclude trial*sex	UN	54	20.02	.005	7	Exclude trial*sex	7539.85	7238.71
9	Exclude trial*edu	UN	54	20.87	.004	7	Keep trial*edu	7540.69	7239.56
10	Exclude age	UN	60	10.59	.001	7	Keep age	7575.88	7241.28
11	Same as model 7	AR(1)	34	355.58	<.001	7		7534.27	7723.87
12	Same as model 7	heterogeneous AR(1)	27	502.63	<.001	7		7817.89	7667.32
13	Same as model 7	homogeneous CS	27	512.86	<.001	7		7828.12	7677.55
14	Same as model 7	heterogeneous CS	34	381.46	<.001	7		7560.15	7560.15

Notes: Cov. Structure, covariance structure; BIC, Bayesian Information Criterion; AIC, Akaike Information Criterion; Edu, years of formal education; Sex was coded 0 = male and 1 = female; Edu and age were mean centered; AR(1), first-order autoregressive covariance structure; and CS, compound symmetry.

Table A.2.1. Pearson's correlations between RAVLT trial scores and the demographic variables ($n = 244$).

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	1-5 total	List B	Age	Sex	Edu
Trial 1	1											
Trial 2	.624	1										
Trial 3	.572	.776	1									
Trial 4	.493	.682	.772	1								
Trial 5	.484	.688	.742	.792	1							
Trial 6	.412	.641	.740	.742	.788	1						
Trial 7	.446	.674	.883	.740	.788	.870	1					
1-5 total	.712	.878	.909	.883	.878	.769	.789	1				
List B	.433	.442	.478	.442	.416	.318	.292	.513	1			
Age	-.225	-.311	-.275	-.206	-.246	-.239	-.269	-.290	-.173	1		
Sex	.242	.324	.306	.240	.287	.284	.252	.330	.118	-.122	1	
Edu	.273	.372	.367	.360	.404	.383	.399	.410	.356	-.168	.010	1

Note: All coefficients are zero-order correlations; Edu = years of education; Sex was coded (0 = male, 1 = female); 1-5 total = trials 1-5 total.

Paper 3

Espenes, J., Lorentzen I. M., Eliassen, I.V., Hessen, E., Waterloo, K., ... & Kirsebom, B.E. (In review).

Regression-based normative data for the D-KEFS Color-Word Interference Test in Norwegian adults ages 20 to 85.

Now published in *The Clinical Neuropsychologist*, 2023.

Regression-based normative data for the D-KEFS Color-Word Interference Test in Norwegian adults ages 20 to 85

Jacob Espenes ^{a, b}, Ingrid Myrvoll Lorentzen ^a, Ingvild Vøllo Eliassen ^{c, d}, Erik Hessen ^{c, d}, Knut Waterloo ^{a, b}, Santiago Timón-Reina ^{c, e}, Tormod Fladby ^{c, f}, Kristine B. Walhovd ^{g, h}, Anders M. Fjell ^{g, h} & Bjørn-Eivind Kirsebom ^{b, a}

^a *Department of Psychology, Faculty of Health Sciences, The Arctic University of Norway, Tromsø, Norway*

^b *Department of Neurology, University Hospital of North Norway, Tromsø, Norway*

^c *Department of Neurology, Akershus University Hospital, Lørenskog, Norway*

^d *Department of Psychology, University of Oslo, Oslo, Norway*

^e *Departamento de Inteligencia Artificial, Universidad Nacional de Educación a Distancia, Madrid, Spain*

^f *Institute of Clinical Medicine, Campus Ahus, University of Oslo, Oslo, Norway*

^g *Center for Lifespan Changes in Brain and Cognition, University of Oslo, Norway*

^h *Computational Radiology and Artificial Intelligence, Department of Radiology and Nuclear Medicine, Oslo University Hospital, Norway*

Running head: Norwegian CWIT (Stroop) norms ages 20 to 85

Jacob Espenes

Department of Psychology, Faculty of Health Sciences, The Arctic University of Norway

Hansine Hansens veg 18, 9019 Tromsø, Norway

E-mail: Johan.j.espenes@uit.no

Telephone number: +47 91734744

Word count: 8119

Keywords: Neuropsychological Tests, Norms, Color-Word Interference Test, CWIT, Norway, Cross-cultural neuropsychology, executive function, D-KEFS, Stroop test

Abstract

Objective: The Delis-Kaplan Executive Function System (D-KEFS) Color-Word-Interference Test (CWIT; AKA Stroop test) is a widely used measure of processing speed and executive function. While test materials and instructions have been translated to Norwegian, only American age-adjusted norms from D-KEFS are available in Norway. We here develop norms in a sample of 1011 Norwegians between 20 and 85 years. Furthermore, we provide indexes for stability over time and assess demographic adjustments applying the D-KEFS norms.

Method: Participants were healthy Norwegian adults from the Center for Lifespan Changes in Brain and Cognition (LCBC) cohort ($n = 899$), the Dementia Disease Initiation ($n = 77$), and Oslo MCI ($n = 35$). Using regression-based norming, we estimated linear and non-linear effects of age, years of education, and sex on the CWIT 1-4 subtests. Stability over time was assessed with intraclass correlation coefficients. The normative adjustment of the D-KEFS norms was assessed with linear regression models.

Results: Increasing age was associated with slower time to completion on all CWIT subtests in a non-linear fashion (accelerated lowering of performance with older age). Women performed better on CWIT-1&3 compared to men. Higher education predicted faster completion time on CWIT-3&4. The original age-adjusted norms from D-KEFS did not adjust for sex or education. Furthermore, we observed significant, albeit small effects of age on all CWIT subtests.

Conclusion: We present demographically adjusted regression-based norms and stability indexes for the D-KEFS CWIT subtests. US D-KEFS norms may be inaccurate for Norwegians with high or low educational attainment, especially women.

Introduction

The basic premise of Stroop tests is to measure an individual's ability to suppress a well-learned automatic response (i.e., word reading) in favor of an unfamiliar and incongruent task (i.e., naming the printed ink color of incongruously named color names) (Rabin, Barr, & Burton, 2005; Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2006). Inhibiting the automatic response is demanding, leading to slower speed and lower accuracy on the incongruent task. This discrepancy is referred to as the 'Stroop interference effect'. While the exact nature of the cognitive processes responsible for the Stroop effect is still discussed, the effect is often regarded to measure the ability to inhibit cognitive interference and maintain focused attention (Scarpina & Tagini, 2017). The prefrontal cortex is highly involved when the Stroop test is performed (Duchek et al., 2013; Keifer & Tranel, 2013; Milham et al., 2002; Miller & Cohen, 2001), and clinical studies have shown that the Stroop interference effect is more pronounced in clinical populations, including patients with frontal lobe dysfunctions (Stuss, Floden, Alexander, Levine, & Katz, 2001), anorexia (Ferro et al., 2005), traumatic brain injury (Ben-David, Nguyen, & van Lieshout, 2011), substance use disorders (Streeter et al., 2008), mild cognitive impairment due to Parkinson's disease (Bezdicek et al., 2015) and dementia by various etiologies (Bayard, Erkes, & Moroni, 2011; Clark et al., 2012).

In cognitively healthy adults, previous research has indicated that a higher level of education is related to better test performance (Brugnolo et al., 2016; Ktaiche, Fares, & Abou-Abbas, 2022; Van der Elst et al., 2006). Consistently, young adults perform better compared to elderly (Brugnolo et al., 2016; Zalonis et al., 2009). Regarding sex differences, there are inconsistent findings with some studies reporting slight sex differences in favor of women (Magnusdottir, Haraldsson, & Sigurdsson, 2021; Van der Elst et al., 2006), while others find no significant difference (Brugnolo et al., 2016; Ktaiche et al., 2022). Some have found significant interaction-effects on Stroop paradigms. Van der Elst et al. (2006) reported that age-related decline was stronger for individuals with less education. On the other hand, Magnusdottir et al. (2021) found that individuals with more education exhibited a stronger age-related decline.

The Stroop test exists in several versions such as the Victoria version (Regard, 1981), the Golden version (Scarpina & Tagini, 2017) and the Color-Word Interference Test (CWIT) from the Delis-Kaplan Executive Function System (D-KEFS) (Delis, Kaplan, & Kramer,

2001). All tasks yield variations of the Stroop interference effect but differ in how the main outcomes are measured. The Victoria Version and Golden Version use the number of correct responses in a fixed amount of time as the outcome. In comparison, the CWIT uses time to completion on a fixed number of test items as the main outcome. Furthermore, the CWIT features a unique fourth condition called inhibition/switching, in which participants are asked to alternate between inhibition and reading color-words. This condition may be more challenging than the classic Stroop color-word inhibition task for some individuals (Lippa & Davis, 2010).

A recent review commissioned by the Norwegian Psychologist Association, the Norwegian Directorate of Health, and the Norwegian Institute of Public Health (Ryder, 2021) indicated that the D-KEFS test battery was amongst the most popular tests used by clinicians in Norway. Also, previous studies have indicated that as much as 91% of Norwegian neuropsychologists use a version of the Stroop test (Egeland et al., 2016). Ryder (2021) reports that despite its popularity, Norwegian norms, in addition to validity and reliability measures, are lacking for the D-KEFS battery. The D-KEFS battery was consequently identified as a priority for validation and norming (Ryder, 2021). To our knowledge, there are no norms outside the original American age-adjusted norms presented in the D-KEFS manual by Delis et al. (2001) available for clinicians and researchers in Norway. Thus, the main objective of this study was to investigate the effect of demographic variables on CWIT performance and provide normative data for the D-KEFS CWIT in a Norwegian sample of cognitively healthy adults. Secondly, we assess the normative adjustment of the original age-adjusted norms from D-KEFS in the same sample of cognitively healthy Norwegian adults. Lastly, for a sub-set of the sample with data from one follow-up testing we provide indexes for stability over time on the D-KEFS CWIT.

Methods and materials

Participants

Normative samples

To develop norms on the Color-Word Interference test (CWIT) we included healthy participants from three research projects in Norway: Studies from the center for lifespan changes in brain and cognition (LCBC) ($n = 899$), the dementia disease initiation study (DDI)

($n = 77$), and the Oslo MCI study ($n = 35$). Descriptive statistics from the normative sample is presented in Table 1. Joint exclusion criteria for all studies were severe somatic or psychiatric illnesses that might influence cognitive functioning. All participants underwent an interview screening for current or previous signs of neurological disorders, epilepsy, stroke, and psychiatric disorders. Participants reporting a subjective experience of cognitive decline such as memory complaints were excluded. The Mini Mental State Examination (MMSE) was used for screening purposes to assess global cognitive functioning. Inclusion criterion were ages 20-85. All participants had Norwegian as their native language and almost all participants were of European ethnicity.

The LCBC (A. M. Fjell et al., 2019) is a multi-disciplinary research center based in Oslo, Norway aimed at investigating normal trajectories of brain and cognition across the lifespan. Healthy participants from LCBC were drawn from three longitudinal sub-projects within the LCBC; Neurocognitive development (Tamnes et al., 2013), Neurocognitive plasticity (de Lange, Bråthen, Rohani, Fjell, & Walhovd, 2018), and Biological Predictors of Memory (Storsve et al., 2014). Participants were recruited through newspaper advertisements and through local Universities and workplaces. Most participants from LCBC were screened for brain abnormalities on MRI scans and participants were excluded if scans showed signs of pathology. A subset of the LCBC sample ($n = 335$) had available follow-up examinations (average test-retest interval 3.4 years) on the CWIT, allowing for test-retest analysis to assess the stability of scores over time. All healthy participants in the test-retest sample fulfilled inclusion criteria and none of the exclusion criteria at baseline testing.

DDI is a Norwegian multi-center longitudinal study on early phases of Alzheimer's Disease and other neurodegenerative diseases (Fladby et al., 2017). Inclusion criterion in the DDI study was age 40-80 years. The Oslo MCI study is the predecessor of the ongoing DDI study and followed the same study protocol as DDI. Assessments in Oslo MCI were performed between 2004 and 2012, and in DDI from 2012 to 2022. Healthy controls from DDI and Oslo MCI were either spouses of symptom group participants, volunteers recruited from advertisements in news outlets, or patients recruited at an orthopedic ward.

[**Table 1** *Descriptive statistics*]

Color-word interference test (CWIT) administration procedures

The CWIT consists of four subtasks. CWIT-1 requires color-naming. The participant is asked to verbally identify the color of solid-colored squares from a sheet of paper. The squares are colored red, green, or blue, and are shown in a random order for a total of 50 items. CWIT-2 requires color-reading. In this subtask, participants are shown the color names “red”, “green”, “blue” (in Norwegian “*rød*”, “*grønn*”, “*blå*”) printed in black ink. The participants are asked to read the color names one-by-one. CWIT 3 (inhibition) corresponds to the classic Stroop task, in which color names are printed with incongruent ink (e.g., “red” printed in green ink). Participants are asked to verbally identify the color of the ink, (thus inhibiting the automated response of reading the color name). CWIT-4 is the inhibition/switching condition. Again, color names are printed with incongruent ink, but approximately fifty percent of the items are enclosed within a black frame. The participant is asked to perform the same task as before (i.e., name the printed color of the ink), except for stimuli that are enclosed within the black frames. Here, the participants are instructed to read the color names. For all subtasks, the participant is asked to respond one-by-one, in succession from left to right, as quickly as possible without making errors. All subtasks are preceded by a brief untimed practice trial consisting of a 10-item sample of the pertinent subtest. The stimuli are organized on laminated sheets in A4 size. Items are arranged in 5 rows of 10 items, totaling 50 items for each subtask. Time to completion and errors are recorded. Errors are recorded as either ‘corrected’ or ‘uncorrected’ by the participant. Administration procedures and standardized instructions for all tasks are described in the D-KEFS manual (Delis, 2005; Delis et al., 2001). Standardized commercially available materials for the D-KEFS CWIT in Norwegian were purchased from Pearson Clinical Norway.

Statistical analyses

Regression norming procedure

We first conducted explorative analyses to evaluate CWIT outcomes and relations to demographic variables before fitting normative models. Pearson correlations indicated significant relationships between age, education, and sex with CWIT 1-4 time to completion (Table 2). We then assessed the distributions for each CWIT subtest for normality which indicated significant positive skewness and kurtosis due to slow completion times for a small part of the normative sample. To normalize measures, we transformed CWIT 1-4 outcomes to

a scaled score distribution ($M = 10$, $SD = 3$) similar to Espenes et al. (2020), Kirsebom et al. (2019), and Testa, Winicki, Pearlson, Gordon, and Schretlen (2009). Measures were normalized using the package “CTT” in R (Willse, 2022). Raw scores were transformed to scaled scores by first determining the percentile ranks of raw scores on CWIT 1-4. Then, percentile ranks were converted to scaled scores in the reversed order so that higher scaled scores related to faster completion time. For instance, the 50th percentile corresponds to scaled score 10, and the 99th percentile corresponds to scaled score 17. Raw score to scaled score conversions are shown in Table 3. Univariate analyses showing the relationships between predictors age and years of education on CWIT 1-4 scaled scores are shown in appendix figure A.1 and A.2.

[**Table 2** *Pearson correlations between time to completion on CWIT 1-4 and demographical variables*]

[**Table 3** *Raw score to scaled score conversion on CWIT 1-4*]

To produce the regression-based norms we performed multiple regression analyses on the CWIT 1-4 scaled scores with age, education, and sex as predictors. We also assessed squared and cubic effects, and interaction terms. Education and age were centered around the mean (i.e., years of education – 15.5) and (age – 46.2) to avoid issues with multicollinearity. For the model selection process, we proceeded similarly to Van der Elst et al. (2006). We started with a full model including all terms related to performance on the CWIT subtests based on previous studies and explorative analyses (Table 2). The preliminary full model included age + age² + age³ + sex + education + education² + education³ + age*sex + education*sex + age*education. With the full model as a reference, we hierarchically dropped terms in a stepwise manner, and compared model fit with the simplified model. Models were compared with ANOVAs for total explained variance (R^2), p -values, and the Bayesian information criterion (BIC). The simplified model was preferred if $p = \geq .01$. The simplified model was subsequently used as reference for further simplification using the same alpha level criterion of $\alpha = .01$. Regression models were reduced until the simplified model explained significantly less variance than the reference model (i.e., $p = \leq .01$). Lastly, we attempted to exchange squared terms in the final models with smooth functions using generalized additive models

(GAMs). The model fit of the GAMs were compared to the linear models following the same procedure as described. BIC and ANOVAs favored the linear models with squared terms, and the smooth functions did not improve model fit to a substantial degree. After reaching the model structures with the best fit for CWIT 1-4 subtests (Table 4), we assessed assumptions of normality and heteroscedasticity using plots of standardized predicted scores and standardized residuals (James, Witten, Hastie, & Tibshirani, 2021). Outliers and influential cases were visually assessed using plots of Cook's distance and standardized residuals. Visual inspection revealed no markedly diverging observations, thus no observations were deleted based on statistical criteria. All analyses were conducted using R version 4.2.1 and packages "dplyr" (Wickham, François, Henry, & Müller, 2022), "CTT" (Willse, 2022), "Psych" (Revelle, 2022) and "mgcv" (Wood & Wood).

Testing the equality of age coefficients on CWIT subtests

Adding to the regression analyses described previously, we considered if the effect of age significantly differed on CWIT subtests. For instance, while the effect of age might significantly predict scores on one subtest, and not the other, this does not infer that the effect of age is different on the subtests (Gelman & Stern, 2006). To test the equality of coefficients we fitted multivariate models (seemingly unrelated regressions) reproducing the normative analyses in Table 4 for two subtests at a time. Then, we tested whether the unstandardized beta coefficients from age obtained through this analysis were equivalent in both models using Z-tests (Table A.1). For these analyses we used an alpha level criterion of $\alpha = .01$ to reject the null hypothesis that the difference between the coefficients is zero (i.e., the coefficients are equal). Multivariate models were fitted because this allows for the calculation of standard errors that are adjusted for the covariance between beta coefficients. Analyses were conducted using R studio version 4.2.1 and the package "Systemfit" (Henningesen & Hamann, 2008) and Z-tests were conducted using the package "Multcomp" (Hothorn et al., 2016).

Errors on the Color-Word Interference Test

To provide normative estimates for errors on CWIT-3 and 4 we summarized corrected and uncorrected errors to a total error score. A total of 936 participants had data on errors. Unfortunately, as we did not record errors on CWIT-1 and 2, we do not provide data regarding the distribution of errors on these subtests. Preliminary analyses indicated that errors on CWIT-3 and 4 were zero-inflated and over-dispersed, as most participants did not make any errors during these subtests. Thus, the variables did not follow a normal distribution

suitable for linear regression analysis. We conducted preliminary analyses to investigate if there were linear associations between errors on the CWIT-3 and 4 with age, education, and sex using Spearman's ROH and Mann-Whitney U tests (Table 2). Analyses were done to assess the need for demographic adjustment or stratification for error measures. Results from these analyses indicated a weak association between errors on CWIT-3 and 4 with demographic variables. We therefore provide percentiles based on the inverse cumulative distribution for errors based on the entire normative sample unstratified according to demographic variables (i.e., unadjusted for age, education, or sex). We then dichotomized the sample into participants who performed 0 errors and ≥ 4 errors on either CWIT-3 or 4 to see if these groups might differ in years of education, age, or sex. In total, 14.1% of the sample made ≥ 4 errors on *either* the CWIT-3 or 4. Thus, ≥ 4 errors on either CWIT-3 or 4 corresponded to a 'low average' score according to neuropsychological nomenclature (Guilmette et al., 2020). We then assessed whether errors on CWIT-3 and 4 were related to performance on the task. First, we compared completion time on CWIT-3 and 4 between individuals who made ≥ 4 errors and individuals who made 0 errors using two-tailed independent samples t-tests without assumptions of equal variance. Further, we correlated errors on CWIT-3 and 4 with time to completion on CWIT-3 and 4 to check for a linear relationship between errors and task performance.

Calculating normative performance using regression-based norms

To determine the normative performance for a given individual (i) on a given test (j), we first calculate the predicted scaled score using the regression equations presented in Table 4. These equations utilize the following formula: Let D be a set of demographic predictors, where d_n represents the n-th element of D; Predicted scaled score_{ij} = intercept_j + sum(beta_coefficient_{dj} * d_{ni}). Then, the individual's raw score on the CWIT is converted to a scaled score using the raw score to scaled score conversion in Table 3. This reflects the individual's obtained scaled score. Lastly, the Z-score of individual (i) on test (j) is computed by $[Z_{ij} = (\text{obtained scaled score}_{ij} - \text{predicted score}_{ij}) / \text{standard deviation of the residual}_j]$, which can be further converted to a T-score by $[(Z_{ij} * 10) + 50]$.

Assessing established American norms from D-KEFS in the Norwegian sample

We computed T-scores based on the original age-adjusted norms from the D-KEFS manual (Delis et al., 2001) on CWIT 1-4 for all participants ($n = 1011$). This resulted in four T-scores

for each participant; T -score on the CWIT-1; CWIT-2; CWIT-3; CWIT-4. To assess if the original age-adjusted norms from D-KEFS sufficiently adjusted for demographical variables in the Norwegian sample, we performed multiple regression analyses with CWIT 1-4 T -scores as dependent variables. Age, years of education, and sex were used as predictors for all analyses. A significant beta-coefficient from any predictor was interpreted as a mal-adjustment in the norms. For these analyses we used a conventional alpha level criterion of $\alpha = .05$. For example, if years of education significantly explained variance in the T -scores, this was interpreted as if the norms did not adequately correct for this demographic variable. Non-significant results were interpreted as an adequate adjustment. T -scores using the new Norwegian norms were calculated for all participants following the procedures detailed in the previous section. We then compared mean T -scores for all participants on the CWIT 1-4 using both the norms from D-KEFS and the new Norwegian norms. Mean T -scores on the CWIT 1-4 were compared using paired samples T -tests without the assumption of equal variances (Table 7). Plots of T -scores on CWIT 1-4 with fitted regression lines for the new Norwegian norms, the D-KEFS norms, and unadjusted T -scores are compared in Figure 1. Lastly, we compared the observed rate of participants scoring below a conventional cut-off (1.5 SD below the normative mean; T -score < 35) on CWIT 1-4 applying the original age-adjusted norms from D-KEFS and the Norwegian norms. Because the T -scores are expected to approximate a normal distribution we used two-tailed one proportion Z -tests to compare the observed rate in the samples with the expected base rate in a theoretical normal distribution (6.7%). The Z -test estimates the probability that the observed sample proportion is equal to the theoretical proportion in the population. For these tests we computed the 99% confidence interval around the sample proportion thereby using a significance level of $\alpha = .01$ (Figure 2). To test if there were significant differences in proportions between the Norwegian norms and the original age-adjusted D-KEFS norms we used paired-samples proportion tests (asymptotic McNemar test without Continuity Correction) (Fagerland, Lydersen, & Laake, 2014).

Norm calculator

To make regression-norms available and easy to use, we provide a free web-based tool that computes the regression equations and provide demographically adjusted T -scores for all CWIT subtests. The tool will be implemented as a self-contained HTML/Javascript webpage but is temporarily available at (<https://contattafiles.s3.us-west->

1.amazonaws.com/tnt30503/ACkqU46CjUb0rss/cwit-calc.html) and is released as open source at (<https://github.com/DDI-NO/cwit-calc>) under Apache License, version 2.0.

Stability over time on the CWIT

A sub-set of the normative sample ($n = 335$) had available follow-up assessments allowing for test-retest correlations assessing stability over time. The sample consisted of 207 women (62%) and 128 men (38%) with a mean age of 52.6 years ($SD = 18.4$) and 15.6 ($SD = 2.9$) years of education at baseline. To ensure that stability indexes remained unified and relevant for clinical practice, participants tested later than 5 years after follow-up were excluded from the analysis ($n = 22$). Thus, the average time between assessments varied between 1 and 5 years with an average test-retest interval of 3.4 years ($SD = 0.9$). Intraclass correlation (ICC) estimates and 95% CIs were calculated based on a single rating, absolute-agreement two-way random-effects model (ICC 2,1) (Shrout & Fleiss, 1979). We specified *a priori* ranges for stability based on conventional reliability classifications from (Koo & Li, 2016). Values between 0.5–0.75 indicate moderate stability and 0.75–0.9 indicate good stability.

Ethics

The Norwegian Regional committees for medical and health research ethics (REK) approved the projects the current study draws upon. Guidelines in the Helsinki declaration of 1964 (revised 2013) and the Norwegian Health and Research Act were followed. All participants gave written informed consents, and were informed of their right to withdraw, as well as potential risks and rewards involved with participation.

Results

Effect of age, education, and sex on CWIT performance

Higher age was on average related to worse performance on all CWIT measures (Table 4). The effects of age and age² were strongly related to performance. On CWIT-1, CWIT-3 and CWIT-4, age and age² accounted for 9.8% - 21.7% of the variance in scores. However, on CWIT-2, age and age² explained merely 2.2% of the variance in scores. Tests of the equality of coefficients indicated that the effect of age was stronger on the complex trials CWIT-3 and CWIT-4 compared to CWIT-1 and CWIT-2 (Table A.1). Furthermore, the effect of age was significantly weaker on CWIT-2 compared to all other subtests. Figure A.1 shows the linear

and quadratic effect of age on all CWIT subtests in the normative sample between 20 and 85 years.

There was a weak but significant positive relationship between years of education and scores on CWIT-3 ($b = 0.078$, $p = .006$, partial $R^2 = 0.8\%$) and CWIT-4 ($b = 0.098$, $p = <.001$, partial $R^2 = 1.2\%$). However, there were no significant associations between years of education and performance on the basic tasks CWIT-1 and CWIT-2 adjusted for sex and age. The relationship between CWIT scores and years of education is shown in Figure A.2.

Women performed significantly better than men on CWIT-1 and CWIT-3, accounting for 1.9% and 0.7% of the variance in scores, respectively. On average, women attained 0.83 higher scaled scores on CWIT-1, and 0.45 on CWIT-3 (Table 4). There were no significant interactions between sex and age, sex and education, or age and education for any CWIT subtests.

[**Table 4** *Normative regression models for CWIT 1-4 based on 1011 healthy Norwegian adults*]

Calculating normative performance on CWIT-1 using regression-based norms

As an example, suppose that a 70-year-old man with 17 years of education completed the CWIT-1 in 35 seconds. The mean age in the normative group was 46.2 and the mean years of education was 15.5 (Table 1). First, we obtain the relevant coefficients from Table 4. The predicted scaled score is calculated by $[9.863 + ((70 - 46.2) * (-0.049)) + ((70 - 46.2)^2 * -0.001) + (0 * 0.825)]$ which is 8.13. From Table 3 we see that a 35 second completion-time on CWIT-1 equates to a scaled score of 7. Thus, the demographically adjusted Z-score is calculated by $[(7 - 8.13) / 2.775]$ giving a Z-score of -0.41. The Z-score can be further converted to a T-score with a mean of 50 and standard deviation of 10 by $[(-0.41 * 10) + 50] = T 46$.

Errors on CWIT-3 and CWIT-4

As shown in Table 2, there were no significant linear associations between demographic variables and errors on CWIT-3 or CWIT-4. Due to the weak association with demographic variables, we report the cumulative percentiles associated with number of errors based on a

subset of the normative sample ($n = 936$) unstratified for age, sex, or educational attainment (Table 5).

[**Table 5.** *Total errors on CWIT-3 and CWIT-4 in a subset of the normative sample ($n = 935$)*]

More errors on the CWIT were associated with longer time to completion on the CWIT. Pearson correlations indicated a positive linear association between total number of errors on CWIT-3 and time to completion, $r(935) = .28$, 95% CI [.219, .338], $p = <.001$, and total errors on the CWIT-4 and time to completion, $r(935) = .408$, 95% CI [.353, .460], $p = <.001$. To illustrate, we dichotomized participants into two groups with ≥ 4 errors on either CWIT-3 and CWIT-4 indicating a ‘low average’ score ($n = 41$), and participants with 0 errors ($n = 250$). As expected, there were no significant differences between groups in age, years of education, or sex. However, participants who made ≥ 4 errors on either task completed the CWIT-3 9.2 seconds slower ($M = 58.7$, $SD = 18.9$) compared to participants with no errors ($M = 49.5$, $SD = 12.2$), $t(45.6) = -3.01$, $M_{\text{diff}} = -9.2$, $p = .004$. On the CWIT-4, participants with ≥ 4 errors completed the subtest 24.4 seconds slower ($M = 79.7$, $SD = 28.9$) compared to the group with 0 errors ($M = 55.2$, $SD = 13.6$), $t(43.0) = -5.32$, $M_{\text{diff}} = -24.4$, $p = <.001$.

Assessing established norms from D-KEFS in the Norwegian sample

As shown in Table 6, results from multiple regression analysis on T -scores calculated using the original age-adjusted D-KEFS norms indicated significant positive effects of age on all CWIT trials, meaning higher age predicted higher T -scores. As shown previously, women performed better than men on CWIT-1 and 3 in the Norwegian sample. However, the norms from D-KEFS did not account for this sex difference, and on average, women attained 2.3 and 1.4 higher T -scores compared to men on the CWIT-1 and 3 (Table 6). Moreover, there was a significant positive association between years of education and CWIT-3 and CWIT-4 T -scores, where participants with higher levels of education received higher T -scores. The combined effect of demographic variables in the age-adjusted scores were low, ranging from 1.6% to 3.0% explained variance. Nevertheless, there were significant mean differences between the D-KEFS norms and the new Norwegian norms (Table 7). On all trials except CWIT-1, the D-KEFS norms produced too high T -scores compared to the expected mean

value of $T = 50$. On CWIT-2 the average T -score using the D-KEFS norms was 52.1; $T = 54.4$ on CWIT-3; and $T = 53.2$ on the CWIT-4.

When utilizing the original age-adjusted norms from D-KEFS the proportion of participants scoring 1.5 SD or more below the normative mean was significantly different compared to the expected base rate on all CWIT subtests (Figure 2). The Norwegian norms were not significantly different compared to the expected base rate and the 99% CI s contained the expected base rate for all subtests ($p > .01$). Results from paired samples proportion tests showed significant differences between the estimated proportion of participants with scores 1.5 SD or more below the normative mean using the Norwegian norms or the original age-adjusted D-KEFS norms ($p < .001$) (Figure 2).

[Table 6 Results from multiple regression analysis on T-scores calculated with the original D-KEFS norms in the normative group ($n = 1011$)]

[Table 7 Paired sample t-tests between T-scores computed using the Norwegian norms and original age-adjusted norms from D-KEFS]

[Figure 2 Percentage of participants in the Norwegian sample ($n = 1011$) with a score 1.5 SD below the normative mean (T -score < 35) on CWIT 1-4.]

Stability over time on the CWIT

Intra-class correlation coefficients and 95% CI s are shown in Table 8. Based on the *a priori* specified ranges, all analyses indicated moderate to good stability in scores between baseline and follow-up using the Norwegian CWIT norms. Slightly lower estimates were obtained with the original D-KEFS norms.

[Table 8 Intra-class correlations between baseline and follow-up on D-KEFS CWIT subtests based on a sub-set of the normative sample ($n = 335$)]

Discussion

Effects of Demographics on the D-KEFS CWIT

We present normative data for the D-KEFS CWIT based on the performance of 1011 healthy Norwegians between 20 and 85 years of age. All four CWIT test scores were related to linear

and quadratic effects of age, indicating a steepening trend towards lower scores for older participants. Quadratic effects of age have been reported on Stroop tests in similar samples spanning the entire adult range (Ktaiche et al., 2022; Van der Elst et al., 2006), but rarely in samples with more restrictive age spans (Bayard et al., 2011; Bezdicek et al., 2015; Magnusdottir et al., 2021; Seo et al., 2008; Tremblay et al., 2016). Consistent with most studies, we found that the basic subtests CWIT-1 (color naming) and CWIT-2 (word reading) were significantly less influenced by age compared to the complex inhibition trial (CWIT-3) and the inhibition/switching trial (CWIT-4) (Adólfssdóttir et al., 2014; Mitrushina, Boone, Razani, & D'Elia, 2005).

Scores on the CWIT may decline with age due to a general age-related slowing of information processing (Salthouse, 1996) and specific deficits in executive functions like inhibitory control (Hasher & Zacks, 1988). Indeed, Adólfssdóttir, Wollschlaeger, Wehling, and Lundervold (2017) showed that higher age significantly predicted slower time to completion on CWIT-3 and 4 after adjusting for processing speed and performance on CWIT-1 and CWIT-2. In other words, when basic non-executive functions were regressed out, there was still an age effect on both CWIT-3 and CWIT-4, thereby implying that there was a specific factor associated with aging beyond generalized slowing. Delis et al. (2001) published contrast measures in the original D-KEFS norms to isolate executive components on the CWIT. However, these contrasts rely on simple subtraction between individual subtest scores, and it has been suggested that this approach might multiply the measurement errors on each test leading to low reliability (Crawford, Sutherland, & Garthwaite, 2008). Unpublished data from the same Norwegian sample used in this study support this, and we hypothesize that a regression-based approach to isolate executive components could mitigate this problem. We therefore aim to develop norms on CWIT-3 and CWIT-4 adjusted for performance for basic tasks using a regression-based approach and compare test-retest reliability with the original contrast scores from D-KEFS in a separate paper.

Effects of education on CWIT scores

Education was significantly, albeit weakly associated with scores on the CWIT-3 and CWIT-4 but was not significantly associated with scores on CWIT-1 and CWIT-2. This is in line with previous studies, where education has been reported to exert a strong influence on the complex Stroop inhibition trial (Bayard et al., 2011; Bezdicek et al., 2015; Brugnolo et al., 2016; Ktaiche et al., 2022; Magnusdottir et al., 2021; Van der Elst et al., 2006). Education is positively associated with full scale IQ (Ritchie & Tucker-Drob, 2018; Steinberg, Bieliauskas,

Smith, & Ivnik, 2005) which might explain why education was related to performance on the complex trials specifically. Moreover, cognitive reserve (Stern et al., 2023) has commonly been proposed as an explanation for how education is related to scores on Stroop tests (Hankee et al., 2016; Ktaiche et al., 2022; Seo et al., 2008; Zalonis et al., 2009). The cognitive reserve hypothesis suggests that individuals with more cognitive reserve, in part obtained through life experiences such as education and engaging occupations, are more resilient to the effects of age-related and pathological decline in the brain which supports cognitive performance (Ewers, 2020; Stern et al., 2023). Resilience may demonstrate because of increased brain reserve such as increased cortical volume, or brain maintenance meaning resistance to neuropathology over time (Stern et al., 2023). Relating to Stroop tests, Van der Elst et al. (2006) showed that individuals with low educational attainment had an accelerated lowering of performance with age compared to individuals with an average or high level of education. This indicates that the individuals with more education were resilient to age-related brain changes and pathology. However, our results indicated a positive effect of education on CWIT-3 and CWIT-4 scores that was independent of age. Therefore, our result might not be related to increased cognitive reserve. In fact, longitudinal studies on the effect of education on cognitive scores and brain health suggest that increased educational attainment on average relates to higher cortical volume and better cognitive performance in adulthood (i.e., brain reserve). However, education does not influence the rate of cortical atrophy or cognitive decline in healthy individuals (i.e., the slope of decline) (Lövdén, Fratiglioni, Glymour, Lindenberger, & Tucker-Drob, 2020; Nyberg et al., 2021). This could explain why education was only related to CWIT scores on average and did not significantly alter the effect of age on scores, like reported by Van der Elst et al. (2006). A possible explanation could be that educational levels might have been more closely related to general cognitive ability (GCA) in Van der Elst et al. (2006) than in our sample. GCA has been related to increased cortical volume in adulthood as well, but unlike education GCA was also associated with regionally lesser rate of cortical atrophy in healthy adults, which indicates brain maintenance (Walhovd et al., 2022).

Compared with our results, some studies report stronger associations between performance on Stroop tests and education (Hankee et al., 2016; Magnúsdóttir et al., 2021) while others report comparable associations (Bayard et al., 2011; Troyer, Leach, & Strauss, 2006). The weak relationship between education and CWIT scores observed in our study might be influenced

by sample characteristics in the normative sample. In particular, the Norwegian sample comprised individuals with relatively uniform and high educational attainment ($M = 15.5$, $SD = 2.9$). So, it follows that samples with uniform levels of education have reduced variance explained by educational attainment. Furthermore, some studies have indicated that the effect of education on scores could be less impactful for the highly educated (Van der Elst et al., 2006), and that the effect of education on Stroop performance could be diminishing after approximately 12 years (Hankee et al., 2016). Reports from Statistics Norway indicate that the educational level of the adult population is divided into three approximately equal parts (Statistics Norway, 2022); mandatory schooling (10 years education); high school level including trade schools (≤ 13 years); university degrees of various lengths (14+ years). Thus, the sample in this study had higher educational attainment than the population average, which may have influenced the relatively weak effect of education on CWIT scores. However, the education range in our sample was 7 to 23 years, and pertinent educational effects on test performance are modelled in our norms at both lower and higher levels of education. The discrepancy between norms is difficult to pinpoint as it could be influenced by several other factors, including the normative estimation method and a variety of cultural influences like educational quality and availability. Regardless, differences between norms highlight the importance that the normative sample resemble the intended population in terms of sample characteristics and geography (Heaton, Avitable, Grant, & Matthews, 1999; International Test Commission, 2001). Specifically, using estimates from foreign samples exhibiting strong effects of education (e.g., Peña-Casanova et al., 2009; Seo et al., 2008) would likely provide inaccurate estimates of performance in the Norwegian sample where education evidently is not as relevant for predicting performance on the CWIT.

Sex differences

Women performed significantly better than men on CWIT-1 (color-naming) and CWIT-3 (inhibition). Previous studies on various Stroop paradigms report inconsistent results regarding sex differences with some studies reporting significant sex-differences (Magnusdottir et al., 2021; Mitrushina et al., 2005; Seo et al., 2008; Tremblay et al., 2016; Van der Elst et al., 2006) while others do not (Adólfssdóttir et al., 2017; Bayard et al., 2011; Hankee et al., 2016; Zalonis et al., 2009). Despite this, any observed difference has consistently favored women. Therefore, it is likely that the effect is small and that a large sample size is needed to detect a sex difference on Stroop tests. A recent article by Sjöberg,

Wilner, D'Souza, and Cole (2023) proposed that the female advantage on Stroop paradigms is related to superior color-naming abilities likely attributed to several specific verbal abilities relevant to performance on the task. These include increased speed on color labelling tasks and better performance on distractor suppression tasks. For a full review, please see Sjoberg et al. (2023). This could explain why we only found a female advantage on CWIT-1 (color naming) and CWIT-3 (inhibition), which has more color stimuli than CWIT-2 (word reading) and CWIT-4 (inhibition/switching).

Errors on CWIT-3 and CWIT-4

Number of errors on the CWIT were not related to age, education, or sex, which is surprising considering existing literature that report significant effects (Tremblay et al., 2016; Troyer et al., 2007; Zalonis et al., 2009; Van der elst et al., 2006). Hanke et al. (2016) report that participants who made errors were significantly older and had less education compared to those with 0 errors. The present study did not find demographic differences between individuals with ≥ 4 errors compared to those with 0 errors. However, consistent with previous studies, our results indicate that errors were significantly related to worse performance on the task. On average, errors on the CWIT-3 and CWIT-4 were correlated with increased time to completion, and participants with ≥ 4 errors completed the CWIT-3 and CWIT-4 significantly slower. For clinical decision making, ≤ 3 errors on CWIT-3 and CWIT-4 should be considered the lower boundary for normal performance corresponding to the ~ 11 - 13^{th} percentile (Table 5). Unfortunately, we do not provide normative estimates for errors on CWIT-1 and CWIT-2. Previous studies indicate that about one in 20 healthy participants make one error on the CWIT-1 or CWIT-2 (Bayard et al., 2011) and multiple errors on these subtests may therefore indicate issues concerning the validity of the test performance. For normative estimates on CWIT-1 and 2 we refer to the original D-KEFS norms by Delis et al. (2001).

Assessment of the original age-adjusted norms from D-KEFS

A key aim of this study was to assess the adequacy of the original age-adjusted norms from D-KEFS in our sample of healthy Norwegians ($n = 1011$). Higher age significantly predicted higher T -scores calculated using norms from D-KEFS. From Figure 1 we can see that the yellow line is above the reference line for $T = 50$ which means that the participants on average performed better than the normative mean from D-KEFS given their age. This indicates that the original norms from D-KEFS slightly exaggerated the detrimental effects of aging on

CWIT performance in the Norwegian sample. As a result, the older participants in the Norwegian sample received slightly elevated *T*-scores on average. Previous studies have found that age-related decline on cognitive tests largely dissipate when adjusting for cerebrovascular pathology, degeneration of structural and functional brain connections, and other pathologies (Anders M. Fjell, Sneve, Grydeland, Storsve, & Walhovd, 2016; Borghesani et al., 2013; Borland, Stomrud, van Westen, Hansson, & Palmqvist, 2020; Harrington et al., 2018; Yu et al., 2015). Age-related decline on the CWIT could therefore be influenced by sub-clinical pathology. Notably, such sub-clinical pathology may be regarded as normal, since most studies with normal healthy participants screened for various pathological conditions indeed report a strong influence of age on scores from Stroop paradigms and other neuropsychological tests (Mitrushina et al., 2005). However, the extent may vary between cohorts. As a result, the comparatively weaker age-effect observed in the Norwegian sample could be due to the Norwegian sample being healthier. These potential differences could be cultural, such as differences in lifestyle and access to health care, or simply cohort-specific, such as cerebrovascular disease prevalence in the study sample. For instance, the Norwegian sample consisted of predominately highly educated individuals that were thoroughly screened which may have caused an over-representation of protective factors in the sample.

The difference between norms may not only be due to cultural differences as cohort differences are observed within cultures as well (Trahan, Stuebing, Fletcher, & Hiscock, 2014). While data regarding Stroop tests specifically is scarce, the literature on other cognitive tests suggests that average cognitive functioning in today's elderly is improved compared to the elderly 20 years ago (Hessel, Kinge, Skirbekk, & Staudinger, 2018; Skirbekk, Stonawski, Bonsang, & Staudinger, 2013). For younger individuals it is less clear with some studies showing that today's young may perform similarly or worse (Bratsberg & Rogeberg, 2018). The improvement of newer cohorts over older cohorts is called the Flynn-effect, which stipulates that improvements in nutrition, educational attainment and quality, health care, health promoting activities such as exercise, and reduction in cardiovascular disease cause newer cohorts to perform better on a variety of cognitive task (Skirbekk et al., 2013). Thus, the disparity between the Norwegian norms and the original age-adjusted norms from D-KEFS published in 2001 may also be due to time of measurement.

[Figure 1 *Plots of T-scores on CWIT 1-4 calculated applying norms from D-KEFS, Norwegian norms and T-scores unadjusted for demographic variables]*

Unsurprisingly, the original age-adjusted norms from D-KEFS did not account for the difference between individuals with high or low educational attainment or the female advantage we observed in the Norwegian sample. As a result, the norms from D-KEFS on average produced higher than expected *T*-scores on the CWIT-2 (2.1 *T*-scores), CWIT-3 (4.4 *T*-scores), and the CWIT-4 (3.2 *T*-scores) compared to the expected value of $T = 50$. As shown in Figure 1, the difference between norms is most apparent in the end ranges of the predictors. Using the D-KEFS norms could have implications for the accurate assessment of very old or very young participants, or individuals with either very high or very low educational attainment. To illustrate, an 80-year-old woman enrolled in this study reported 17 years of education and performed the CWIT-3 in 78 seconds and the CWIT-4 in 85 seconds. According to the Norwegian norms, her scores equate to $T = 43$ on CWIT-3, and $T = 47$ on CWIT-4, thus reflecting a below average performance. Using the D-KEFS norms her scores were $T = 57$ on both tasks, i.e., 1.4 *SD* and 1 *SD* higher compared to the Norwegian norms.

We found that the proportion of participants scoring below a conventional cut-off set at 1.5 *SD* below the normative mean significantly differed from the expected proportion when using the original age-adjusted norms from D-KEFS. From Figure 2 it is apparent that the D-KEFS norms located fewer-than-expected participants with low scores on all CWIT subtests. Furthermore, the percentage of participants with low scores significantly differed between the norms with the D-KEFS norms identifying significantly fewer participants ($p < .001$). This indicates that the norms from D-KEFS have a lower sensitivity for correctly identifying individuals with low scores on the CWIT in the Norwegian sample. Although not statistically significant, the Norwegian norms located more participants with low scores on CWIT-2 and CWIT-4 than we expected (8.6% and 8.3% respectively). The Norwegian norms were expected to match the theoretical base rate of 6.7% more closely since the norms were produced in the same sample and scores were transformed to follow a normal distribution. The difference is likely caused by some skewness in the CWIT-2 and 4 scaled scores despite the normalization procedures which caused a slight over representation of participants around this cut-off. Future studies should assess the Norwegian norms in an independent sample of Norwegians to address whether the new norms equal the theoretical base rate of impairment.

Correlations between baseline performance and follow-up on the CWIT

All psychological tests should have available evidence of reliability that is relevant to the intended population (International test commission, 2001). Ryder (2021) identified that tests from the D-KEFS battery were lacking reliability estimates based on a Norwegian sample. In this study we had test-retest scores based on a relatively long follow-up ($M = 3.4$ years), and test-retest correlations are therefore assumed to not just be a measure of reliability, but also reflect true change rates with age. For instance, a low correlation would typically be interpreted as low reliability, but it could also mean that some participants have a different slope (i.e., change rate) from baseline to follow-up. We therefore characterize the test-retest correlations as stability of scores over time. The results indicated moderate to good correlations for all measures, with slightly better correlation for the complex trials CWIT-3 and CWIT-4 (Table 8). Using the Norwegian norms resulted in marginally better correlation compared to the original D-KEFS norms, likely due to the slight mal-adjustments in age, education, and sex previously reported. The difference between coefficients using Norwegians norms and D-KEFS norms were not tested, although the 95% *CI* overlapped and the difference in coefficients would likely not fulfill conventional criteria for statistical significance.

Limitations

The current study is subject to some limitations. Firstly, neuropsychological norms are typically intended to give an estimate of an individual's score compared to a broad target population, e.g., healthy Norwegians between 20 and 85 years old. The representativeness of a normative sample is therefore crucial for the accuracy of the normative estimates. Most of the participants included in this study were recruited as healthy participants from advertisements, university, and workplaces, and could be susceptible to biases associated with convenience sampling methods. That is, the sample estimates may not generalize to the broad target population due to unknown biases arising from a non-probability sampling method (Jager, Putnick, & Bornstein, 2017). Relatedly, most of the sample were native Norwegian speakers of European ethnicity which does not reflect the multicultural landscape in Norway. The norms are likely less accurate for people with Norwegian as a second language and immigrants. Despite this, as the first normative study outside the original age-adjusted norms presented in the D-KEFS manual (Delis et al., 2001) we believe our norms contribute to an improvement in the accuracy of CWIT assessments in Norway. Another limitation of this study is the lack of participants in the middle-age. However, the norms rely on the joint

estimation of the average effect across the included age span to calculate predicted scores and deviation from the predicted scores. Thus, it is unlikely that the lack of participants in the middle-age greatly affect the norms' ability to predict scores for individuals in this age range. Unfortunately, we were not able to source an independent sample to compare the new Norwegian norms with the original age-adjusted norms from D-KEFS. Instead, we assessed the adequacy of the original D-KEFS norms in our normative sample. Future studies should assess the validity of both the new norms and the original D-KEFS norms in an independent sample of Norwegians. Lastly, we did not formally screen for visual impairment but relied on self-report of visual deficits. Though participants used glasses when applicable, we cannot guarantee that undiagnosed visual impairment did not influence some participants' scores.

Conclusion

We propose regression-based norms for the Delis-Kaplan Color Word Interference Test (CWIT) based on a sample of healthy Norwegian adults between 20 and 85 years old ($n = 1011$). As far as we know, this is the first published study providing norms on the D-KEFS CWIT apart from the original age-adjusted norms from D-KEFS (Delis et al., 2001). Our results indicate that lower age, higher education, and female sex significantly predicted improved performance on the CWIT. The original age-adjusted norms from D-KEFS overestimated the difference between young and old participants and did not adjust for the female advantage or effects of education in the Norwegian sample. Consequently, normative estimates from the original D-KEFS norms may be inaccurate for young or old individuals with either low or high educational attainment. The norms from D-KEFS identified significantly fewer-than-expected participants with low scores on CWIT 1-4 in the Norwegian sample. Low scores were defined as scores 1.5 *SD* or more below the normative mean. Thus, the D-KEFS norms had a lower sensitivity for detecting individuals with potential executive deficits compared to the Norwegian norms. In the Norwegian sample, ≥ 4 errors on CWIT-3 and CWIT-4 corresponded to the $\sim 5^{\text{th}}$ percentile, indicative of a borderline impaired performance. Errors were unrelated to demographical variables, but increased number of errors were significantly related to slower time to completion on the CWIT-3 and CWIT-4. The CWIT showed moderate to good test-retest stability in the Norwegian sample with a 3.4-year average follow-up time. For ease of use and quick computation of the norms we provide a normative calculator available at (<https://contattafiles.s3.us-west-1.amazonaws.com/tnt30503/ACkqU46CjUb0rss/cwit-calc.html>).

Funding

This work was supported by the University of Tromsø - the Arctic University of Norway; the university of Oslo; the Norwegian Research Council, Helse Nord under grant number HNF1540-20. Additional support was received from ERC grants (283634 and 725025 to AMF), Research Council of Norway grants to KBW and AMF, and European Commission (EC) EU Horizon 2020 Grant agreement number 732592 (Lifebrain project). The funding sources were not involved in the drafting of this manuscript.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy of the research participants. The LCBC, DDI and Oslo MCI datasets has restricted access, requests can be made to the corresponding author, and some of the data can be made available given appropriate ethical and data protection approvals.

Acknowledgments

We thank all participants and collaborators at the LCBC and DDI study sites for contributing to data used in this study.

Conflict of Interest

The authors have no conflict of interest to report.

Appendix

[Figure A.1 *Scatterplots of age² and CWIT 1-4 scaled scores in the normative sample (n = 1011).]*

[Figure A.2 *Scatterplots of years of education and CWIT 1-4 scaled scores in the normative sample (n = 1011).]*

[Table A.1 Equality of age coefficients on CWIT subtests]

References

- Adólfssdóttir, S., Haász, J., Wehling, E., Ystad, M., Lundervold, A., & Lundervold, A. J. (2014). Salient measures of inhibition and switching are associated with frontal lobe gray matter volume in healthy middle-aged and older adults. *Neuropsychology, 28*(6), 859.
- Adólfssdóttir, S., Wollschlaeger, D., Wehling, E., & Lundervold, A. J. (2017). Inhibition and switching in healthy aging: a longitudinal study. *Journal of the International Neuropsychological Society, 23*(1), 90-97.
- Bayard, S., Erkes, J., & Moroni, C. (2011). Victoria Stroop Test: normative data in a sample group of older people and the study of their clinical applications in the assessment of inhibition in Alzheimer's disease. *Archives of Clinical Neuropsychology, 26*(7), 653-661.
- Beck, A. T., Steer, R. A., & Brown, G. K. (1987). *Beck depression inventory*: Harcourt Brace Jovanovich New York:.
- Ben-David, B. M., Nguyen, L. L., & van Lieshout, P. H. (2011). Stroop effects in persons with traumatic brain injury: selective attention, speed of processing, or color-naming? A meta-analysis. *J Int Neuropsychol Soc, 17*(2), 354-363. doi:10.1017/s135561771000175x
- Bezdicek, O., Lukavsky, J., Stepankova, H., Nikolai, T., Axelrod, B. N., Michalec, J., . . . Kopecek, M. (2015). The Prague Stroop Test: Normative standards in older Czech adults and discriminative validity for mild cognitive impairment in Parkinson's disease. *Journal of clinical and experimental neuropsychology, 37*(8), 794-807.
- Borghesani, P. R., Madhyastha, T. M., Aylward, E. H., Reiter, M. A., Swarny, B. R., Schaie, K. W., & Willis, S. L. (2013). The association between higher order abilities, processing speed, and age are variably mediated by white matter integrity during typical aging. *Neuropsychologia, 51*(8), 1435-1444.
- Borland, E., Stomrud, E., van Westen, D., Hansson, O., & Palmqvist, S. (2020). The age-related effect on cognitive performance in cognitively healthy elderly is mainly caused by underlying AD pathology or cerebrovascular lesions: implications for cutoffs regarding cognitive impairment. *Alzheimer's research & therapy, 12*(1), 1-10.
- Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. *Proc Natl Acad Sci U S A, 115*(26), 6674-6678. doi:10.1073/pnas.1718793115
- Brugnolo, A., De Carli, F., Accardo, J., Amore, M., Bosia, L., Bruzzaniti, C., . . . Ferrara, M. (2016). An updated Italian normative dataset for the Stroop color word test (SCWT). *Neurological Sciences, 37*(3), 365-372.
- Clark, L. R., Schiehser, D. M., Weissberger, G. H., Salmon, D. P., Delis, D. C., & Bondi, M. W. (2012). Specific measures of executive function predict cognitive decline in older adults. *J Int Neuropsychol Soc, 18*(1), 118-127. doi:10.1017/s1355617711001524

- Crawford, J. R., Sutherland, D., & Garthwaite, P. H. (2008). On the reliability and standard errors of measurement of contrast measures from the D-KEFS. *Journal of the International Neuropsychological Society*, *14*(6), 1069-1073.
- de Lange, A.-M. G., Bråthen, A. C. S., Rohani, D. A., Fjell, A. M., & Walhovd, K. B. (2018). The temporal dynamics of brain plasticity in aging. *Cerebral cortex*, *28*(5), 1857-1865.
- Delis, D. (2005). Delis-Kaplan executive function system, Norwegian version. *Bromma: Pearson Assessment*.
- Delis, D., Kaplan, E., & Kramer, J. (2001). D-KEFS: examiners manual. *San Antonio, TX: The Psychological Corporation*.
- Duchek, J. M., Balota, D. A., Thomas, J. B., Snyder, A. Z., Rich, P., Benzinger, T. L., . . . Ances, B. M. (2013). Relationship Between Stroop Performance and Resting State Functional Connectivity in Cognitively Normal Older Adults. *Neuropsychology*, *27*(5), 516-528. doi:10.1037/a0033402
- Egeland, J., Løvstad, M., Norup, A., Nybo, T., Persson, B. A., Rivera, D. F., . . . Arango-Lasprilla, J. C. (2016). Following international trends while subject to past traditions: neuropsychological test use in the Nordic countries. *The Clinical Neuropsychologist*, *30*(sup1), 1479-1500. Retrieved from <https://doi.org/10.1080/13854046.2016.1237675>. doi:10.1080/13854046.2016.1237675
- Espenes, J., Hessen, E., Eliassen, I. V., Waterloo, K., Eckerström, M., Sando, S. B., . . . Kirsebom, B. E. (2020). Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years. *Clin Neuropsychol*, *34*(sup1), 110-126. doi:10.1080/13854046.2020.1829068
- Ewers, M. (2020). Reserve in Alzheimer's disease: update on the concept, functional mechanisms and sex differences. *Current opinion in psychiatry*, *33*(2), 178-184.
- Fagerland, M. W., Lydersen, S., & Laake, P. (2014). Recommended tests and confidence intervals for paired binomial proportions. *Statist. Med*, *33*(16), 2850-2875. doi:10.1002/sim.6148
- Ferro, A. M., Brugnolo, A., De Leo, C., Dessi, B., Girtler, N., Morbelli, S., . . . Rodriguez, G. (2005). Stroop interference task and single-photon emission tomography in anorexia: A preliminary report. *International Journal of Eating Disorders*, *38*(4), 323-329. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/eat.20203>. doi:https://doi.org/10.1002/eat.20203
- Fladby, T., Pålhaugen, L., Selnes, P., Bråthen, G., Hessen, E., Almdahl, I. S., . . . Espenes, R. (2017). Detecting at-risk Alzheimer's disease cases. *Journal of Alzheimer's Disease*, *60*(1), 97-105.
- Fjell, A. M., Chen, C. H., Sederevicius, D., Sneve, M. H., Grydeland, H., Krogsrud, S. K., . . . Walhovd, K. B. (2019). Continuity and Discontinuity in Human Cortical Development and Change From Embryonic Stages to Old Age. *Cereb Cortex*, *29*(9), 3879-3890. doi:10.1093/cercor/bhy266
- Fjell, A. M., Sneve, M. H., Grydeland, H., Storsve, A. B., & Walhovd, K. B. (2016). The Disconnected Brain and Executive Function Decline in Aging. *Cerebral cortex*, *27*(3), 2303-2317. Retrieved from <https://doi.org/10.1093/cercor/bhw082>. doi:10.1093/cercor/bhw082
- 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*(3), 189-198.
- Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American statistician*, *60*(4), 328-331. doi:10.1198/000313006X152649
- Guilmette, T. J., Sweet, J. J., Hebben, N., Koltai, D., Mahone, E. M., Spiegler, B. J., . . . Westerveld, M. (2020). American Academy of Clinical Neuropsychology consensus conference statement on uniform labeling of performance test scores. *The Clinical Neuropsychologist*, *34*(3), 437-453. Retrieved from <https://doi.org/10.1080/13854046.2020.1722244>. doi:10.1080/13854046.2020.1722244
- Hankee, L. D., Preis, S. R., Piers, R. J., Beiser, A. S., Devine, S. A., Liu, Y., . . . Au, R. (2016). Population Normative Data for the CERAD Word List and Victoria Stroop Test in Younger- and Middle-Aged Adults: Cross-Sectional Analyses from the Framingham Heart Study. *Experimental Aging*

- Research*, 42(4), 315-328. Retrieved from <https://doi.org/10.1080/0361073X.2016.1191838>. doi:10.1080/0361073X.2016.1191838
- Harrington, K. D., Schembri, A., Lim, Y. Y., Dang, C., Ames, D., Hassenstab, J., . . . Rowe, C. C. (2018). Estimates of age-related memory decline are inflated by unrecognized Alzheimer's disease. *Neurobiology of Aging*, 70, 170-179.
- Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. *Psychology of learning and motivation*, 22, 193-225.
- Heaton, R. K., Avitable, N., Grant, I., & 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(4), 572-582.
- Henningsen, A., & Hamann, J. D. (2008). systemfit: A package for estimating systems of simultaneous equations in R. *Journal of statistical software*, 23, 1-40.
- Hessel, P., Kinge, J. M., Skirbekk, V., & Staudinger, U. M. (2018). Trends and determinants of the Flynn effect in cognitive functioning among older individuals in 10 European countries. *Journal of Epidemiology and Community Health*, 72(5), 383-389. Retrieved from <https://jech.bmj.com/content/jech/72/5/383.full.pdf>. doi:10.1136/jech-2017-209979
- Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., Scheibe, S., & Hothorn, M. T. (2016). Package 'multcomp'. *Simultaneous inference in general parametric models. Project for Statistical Computing, Vienna, Austria*.
- International Test, C. (2001). International Guidelines for Test Use. *International Journal of Testing*, 1(2), 93-114. Retrieved from https://doi.org/10.1207/S15327574IJT0102_1. doi:10.1207/S15327574IJT0102_1
- Jager, J., Putnick, D. L., & Bornstein, M. H. (2017). II. MORE THAN JUST CONVENIENT: THE SCIENTIFIC MERITS OF HOMOGENEOUS CONVENIENCE SAMPLES. *Monogr Soc Res Child Dev*, 82(2), 13-30. doi:10.1111/mono.12296
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. New York, NY: New York, NY: Springer.
- Keifer, E., & Tranel, D. (2013). A neuropsychological investigation of the Delis-Kaplan executive function system. *Journal of clinical and experimental neuropsychology*, 35(10), 1048-1059.
- Kirsebom, B.-E., Espenes, R., Hessen, E., Waterloo, K., Johnsen, S. H., Gundersen, E., . . . Fladby, T. (2019). Demographically adjusted CERAD wordlist test norms in a Norwegian sample from 40 to 80 years. *The Clinical Neuropsychologist*, 33(sup1), 27-39. Retrieved from <https://doi.org/10.1080/13854046.2019.1574902>. doi:10.1080/13854046.2019.1574902
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2), 155-163.
- Krogsrud, S. K., Mowinckel, A. M., Sederevicius, D., Vidal-Piñeiro, D., Amlien, I. K., Wang, Y., . . . Fjell, A. M. (2021). Relationships between apparent cortical thickness and working memory across the lifespan - Effects of genetics and socioeconomic status. *Developmental Cognitive Neuroscience*, 51, 100997. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1878929321000876>. doi:<https://doi.org/10.1016/j.dcn.2021.100997>
- Ktaiche, M., Fares, Y., & Abou-Abbas, L. (2022). Stroop color and word test (SCWT): Normative data for the Lebanese adult population. *Applied Neuropsychology: Adult*, 29(6), 1578-1586.
- Lippa, S. M., & Davis, R. N. (2010). Inhibition/switching is not necessarily harder than inhibition: An analysis of the D-KEFS Color-Word Interference Test. *Archives of Clinical Neuropsychology*, 25, 146-152. doi:10.1093/arclin/acq001
- Lövdén, M., Fratiglioni, L., Glymour, M. M., Lindenberger, U., & Tucker-Drob, E. M. (2020). Education and Cognitive Functioning Across the Life Span. *Psychol Sci Public Interest*, 21(1), 6-41. doi:10.1177/1529100620920576
- Magnus, P., Irgens, L. M., Haug, K., Nystad, W., Skjærven, R., & Stoltenberg, C. (2006). Cohort profile: the Norwegian mother and child cohort study (MoBa). *International journal of epidemiology*, 35(5), 1146-1150.

- Magnusdottir, B., Haraldsson, H., & Sigurdsson, E. (2021). Trail making test, stroop, and verbal fluency: Regression-based norms for the icelandic population. *Archives of Clinical Neuropsychology*, *36*(2), 253-266.
- Milham, M. P., Erickson, K. I., Banich, M. T., Kramer, A. F., Webb, A., Wszalek, T., & Cohen, N. J. (2002). Attentional Control in the Aging Brain: Insights from an fMRI Study of the Stroop Task. *Brain and Cognition*, *49*(3), 277-296. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0278262601915015>. doi:<https://doi.org/10.1006/brcg.2001.1501>
- Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167-202. Retrieved from <https://www.annualreviews.org/doi/abs/10.1146/annurev.neuro.24.1.167>. doi:[10.1146/annurev.neuro.24.1.167](https://doi.org/10.1146/annurev.neuro.24.1.167)
- Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment*: Oxford University Press.
- Nyberg, L., Magnussen, F., Lundquist, A., Baaré, W., Bartrés-Faz, D., Bertram, L., . . . Fjell, A. M. (2021). Educational attainment does not influence brain aging. *Proc Natl Acad Sci U S A*, *118*(18). doi:[10.1073/pnas.2101644118](https://doi.org/10.1073/pnas.2101644118)
- Peña-Casanova, J., Quiñones-Úbeda, S., Gramunt-Fombuena, N., Quintana, M., Aguilar, M., Molinuevo, J. L., . . . Team, f. t. N. S. (2009). Spanish Multicenter Normative Studies (NEURONORMA Project): Norms for the Stroop Color-Word Interference Test and the Tower of London-Drexel. *Archives of Clinical Neuropsychology*, *24*(4), 413-429. Retrieved from <https://doi.org/10.1093/arclin/acp043>. doi:[10.1093/arclin/acp043](https://doi.org/10.1093/arclin/acp043)
- Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: a survey of INS, NAN, and APA Division 40 members. *Arch Clin Neuropsychol*, *20*(1), 33-65. doi:[10.1016/j.acn.2004.02.005](https://doi.org/10.1016/j.acn.2004.02.005)
- Regard, M. (1981). Stroop test–Victoria version. *Victoria, BC: Neuropsychological Laboratory, University of Victoria*.
- Revelle, W. (2022). Psych: procedures for psychological, psychometric, and personality research. 2020. *R package version*, *2*(7).
- Ritchie, S. J., & Tucker-Drob, E. M. (2018). How Much Does Education Improve Intelligence? A Meta-Analysis. *Psychol Sci*, *29*(8), 1358-1369. doi:[10.1177/0956797618774253](https://doi.org/10.1177/0956797618774253)
- Rivera, D., Perrin, P. B., Stevens, L. F., Garza, M. T., Weil, C., Saracho, C. P., . . . Arango-Lasprilla, J. C. (2015). Stroop Color-Word Interference Test: Normative data for the Latin American Spanish speaking adult population. *NeuroRehabilitation*, *37*(4), 591-624. doi:[10.3233/nre-151281](https://doi.org/10.3233/nre-151281)
- Ryder, T. (2021). Testkvalitetsprosjektet-del 1: Norske psykologers testholdninger og testbruk. *Tidsskrift for Norsk psykologforening*, *58*(1), 28-37.
- Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403-428. doi:[10.1037/0033-295X.103.3.403](https://doi.org/10.1037/0033-295X.103.3.403)
- Scarpina, F., & Tagini, S. (2017). The stroop color and word test. *Frontiers in psychology*, *8*, 557.
- Seo, E. H., Lee, D. Y., Choo, I. H., Kim, S. G., Kim, K. W., Youn, J. C., . . . Woo, J. I. (2008). Normative study of the Stroop Color and Word Test in an educationally diverse elderly population. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, *23*(10), 1020-1027.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.
- Sjoberg, E. A., Wilner, R. G., D'Souza, A., & Cole, G. G. (2023). The Stroop Task Sex Difference: Evolved Inhibition or Color Naming? *Archives of Sexual Behavior*, *52*(1), 315-323.
- Skirbekk, V., Stonawski, M., Bonsang, E., & Staudinger, U. M. (2013). The Flynn effect and population aging. *Intelligence (Norwood)*, *41*(3), 169-177. doi:[10.1016/j.intell.2013.02.001](https://doi.org/10.1016/j.intell.2013.02.001)
- Statistics Norway. (2022). Facts about education in Norway 2022 - key figures 2020. Retrieved from <https://www.ssb.no/en/utdanning/utdanningsniva/artikler/facts-about-education-in->

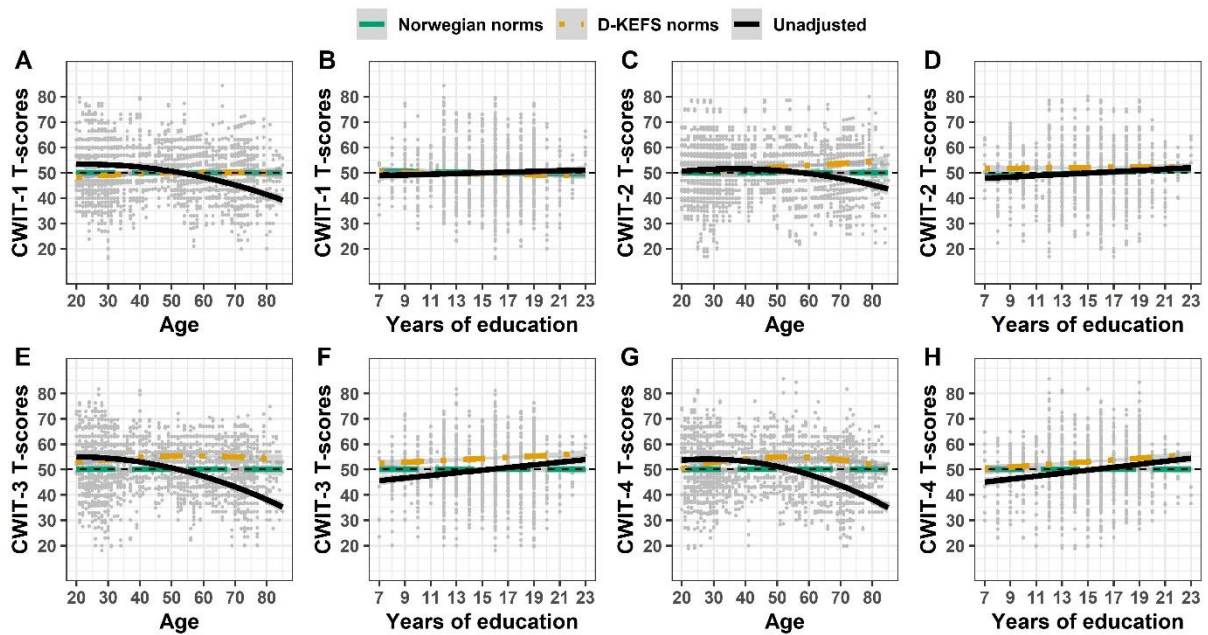
- norway-2022/_/attachment/inline/5ea7e453-8d7c-4423-80fb-b8bd2b8a0340:10c4b259c03df8c44ee58708542c2cf7977a9b4a/FOU-2022-web-en.pdf
- Steinberg, B. A., Bieliauskas, L. A., Smith, G. E., & Ivnik, R. J. (2005). Mayo's older Americans normative studies: Age- and IQ-adjusted norms for the Trailmaking Test, the Stroop Test, and Mae Controlled Oral Word Association Test. *Clinical neuropsychologist*, *19*(3-4), 329-377. doi:10.1080/13854040590945210
- Stern, Y. (2002). What is cognitive reserve? Theory and research application of the reserve concept. *Journal of the International Neuropsychological Society*, *8*(3), 448-460.
- Stern, Y. (2009). Cognitive reserve. *Neuropsychologia*, *47*(10), 2015-2028.
- Stern, Y., Albert, M., Barnes, C. A., Cabeza, R., Pascual-Leone, A., & Rapp, P. R. (2023). A framework for concepts of reserve and resilience in aging. *Neurobiology of Aging*, *124*, 100-103. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0197458022002548>. doi:<https://doi.org/10.1016/j.neurobiolaging.2022.10.015>
- Storsve, A. B., Fjell, A. M., Tamnes, C. K., Westlye, L. T., Overbye, K., Aasland, H. W., & Walhovd, K. B. (2014). Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: regions of accelerating and decelerating change. *Journal of Neuroscience*, *34*(25), 8488-8498.
- Streeter, C. C., Terhune, D. B., Whitfield, T. H., Gruber, S., Sarid-Segal, O., Silveri, M. M., . . . Tian, H. (2008). Performance on the Stroop predicts treatment compliance in cocaine-dependent individuals. *Neuropsychopharmacology*, *33*(4), 827-836.
- Stuss, D. T., Floden, D., Alexander, M. P., Levine, B., & Katz, D. (2001). Stroop performance in focal lesion patients: dissociation of processes and frontal lobe lesion location. *Neuropsychologia*, *39*(8), 771-786. doi:10.1016/s0028-3932(01)00013-6
- Tamnes, C. K., Walhovd, K. B., Dale, A. M., Østby, Y., Grydeland, H., Richardson, G., . . . Due-Tønnessen, P. (2013). Brain development and aging: overlapping and unique patterns of change. *Neuroimage*, *68*, 63-74.
- Testa, S. M., Winicki, J. M., Pearlson, G. D., Gordon, B., & Schretlen, D. J. (2009). Accounting for estimated IQ in neuropsychological test performance with regression-based techniques. *J Int Neuropsychol Soc*, *15*(6), 1012-1022. doi:10.1017/s1355617709990713
- Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn Effect: A Meta-Analysis. *Psychol Bull*, *140*(5), 1332-1360. doi:10.1037/a0037173
- Tremblay, M.-P., Potvin, O., Belleville, S., Bier, N., Gagnon, L., Blanchet, S., . . . Hudon, C. (2016). The victoria stroop test: normative data in Quebec-French adults and elderly. *Archives of Clinical Neuropsychology*, *31*(8), 926-933.
- Troyer, A. K., Leach, L., & Strauss, E. (2006). Aging and response inhibition: Normative data for the Victoria Stroop Test. *Aging, Neuropsychology, and Cognition*, *13*(1), 20-35.
- Van der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., & Jolles, J. (2006). 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*(1), 62-79. doi:10.1177/1073191105283427
- Walhovd, K. B., Nyberg, L., Lindenberger, U., Amlien, I. K., Sørensen, Ø., Wang, Y., . . . Fjell, A. M. (2022). Brain aging differs with cognitive ability regardless of education. *Sci Rep*, *12*(1), 13886. doi:10.1038/s41598-022-17727-6
- Wickham, H., François, R., Henry, L., & Müller, K. (2022). dplyr: A grammar of data manipulation. *R package version 1.0.8*. URL: <https://CRAN.R-project.org/package=dplyr>.
- Willse, J. T. (2022). CTT: Classical Test Theory Functions. R package version 2.3.3. URL: <https://rdrr.io/cran/CTT/>
- Wood, S., & Wood, M. S. Package 'mgcv'.
- Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Leirer, V. O. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of psychiatric research*, *17*(1), 37-49.

Yu, L., Boyle, P. A., Segawa, E., Leurgans, S., Schneider, J. A., Wilson, R. S., & Bennett, D. A. (2015). Residual decline in cognition after adjustment for common neuropathologic conditions. *Neuropsychology, 29*(3), 335.

Zalonis, I., Christidi, F., Bonakis, A., Kararizou, E., Triantafyllou, N. I., Paraskevas, G., . . . Vasilopoulos, D. (2009). The stroop effect in Greek healthy population: normative data for the Stroop Neuropsychological Screening Test. *Archives of Clinical Neuropsychology, 24*(1), 81-88.

Figure 1

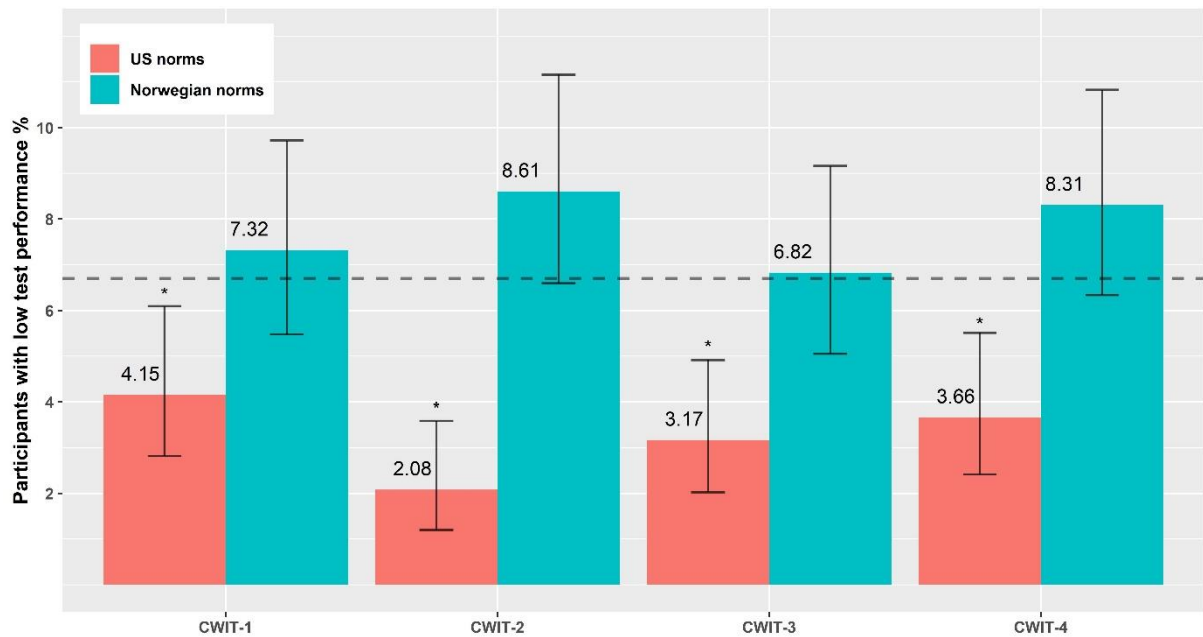
Plots of T-scores on CWIT-1 and CWIT-3 calculated applying norms from D-KEFS, Norwegian norms, and T-scores unadjusted for demographic variables.



Note. Linear regression lines are fitted for years of education and squared lines for age; for all figures a horizontal line from $T = 50$ represents the ideal normative correction and deviation from this line may indicate maladjustment in the norms.

Figure 2

Percentage of participants in the Norwegian sample ($n = 1011$) with a score 1.5 SD below the normative mean (T -score < 35) on CWIT 1-4.



Note. Dotted line indicates the expected base rate for 1.5 SD below the normative mean (6.7%). Error bars indicate the 99% confidence interval (CI) around the estimate. *CI does not contain the expected base rate ($p < .01$). Paired samples proportion tests indicated significant difference between rates from US norms and Norwegian norms on all CWIT subtests ($p < .001$).

Table 1

Descriptive statistics for the normative sample ($n = 1011$).

Variable	Mean (SD)	[Min, Max]	Median
Age	46.2 (19.4)	[20, 85]	43
Female n (%)	675 (66.8%)		
Years of education	15.5 (2.9)	[7, 23]	16
MMSE ¹	29.1 (1.1)	[24, 30]	29
CWIT-1 raw score	30.1 (5.7)	[15, 60]	29
CWIT-2 raw score	22.0 (4.3)	[13, 79]	21
CWIT-3 raw score	53.0 (14.6)	[25, 154]	50
CWIT-4 raw score	60.4 (17.3)	[28, 172]	57
Total errors CWIT-3 ¹	1.0	[0, 11]	1
Total errors CWIT-4 ¹	1.1	[0, 11]	1

Note. SD = standard deviation of the mean; n = count; CWIT = Color-Word Interference Test; Min = lowest score; Max = highest score; MMSE = Mini Mental State Examination; ¹ 76 participants had missing values on errors and 40 on MMSE.

Table 2*Pearson correlation between time to completion on CWIT 1-4 and demographical variables*

Parameter	Age	Age ²	Education	Sex
CWIT-1 raw	.366*	.380*	-.033	-.187*
CWIT-2 raw	.124*	.135*	-.065	-.041
CWIT-3 raw	.500*	.519*	-.150*	-.134*
CWIT-4 raw	.450*	.470*	-.178*	-.084*
Errors CWIT-3 ¹	-.068	-.068	-.051	.017
Errors CWIT-4 ¹	-.014	-.014	-.042	.035

Note. *Statistically significant ($p < .01$); ¹For errors, Spearman's rho is reported for continuous variables. Sex-differences on errors were tested with Mann-Whitney test and the rank-biserial correlation is reported.

Table 3*Raw score to scaled score conversion on CWIT 1-4*

Scaled score	Percentile	CWIT-1	CWIT-2	CWIT-3	CWIT-4
1	0.1	≥56	≥52	≥130	≥168
2	0.4	48-55	37-51	114-129	133-167
3	1	46-47	34-36	99-113	117-132
4	2	42-45	31-33	85-98	101-116
5	5	40-41	28-30	77-84	88-100
6	9	37-39	27	69-76	78-87
7	16	35-36	25-26	63-68	71-77
8	25	33-34	24	58-62	65-70
9	37	31-32	23	53-57	60-64
10	50	29-30	21-22	49-52	55-59
11	63	27-28	20	45-48	51-54
12	75	26	19	42-44	48-50
13	84	25	18	40-41	45-47
14	91	23-24	17	37-39	42-44
15	95	22	16	35-36	39-41
16	98	21	15	33-34	37-38
17	99	20		31-32	34-36
18	99.6	19	≤14	27-30	31-33
19	99.9	≤18		≤26	≤30

Note. Scaled scores are not adjusted for demographical variables and are only used for computing the regression equations in Table 4.

Table 4*Normative regression models for CWIT 1-4 based on 1011 healthy Norwegian adults*

	Parameter	<i>b</i>	<i>b</i> 95 % CI [LL, UL]	<i>s.e.</i>	<i>t</i>	<i>p</i>	Partial <i>R</i> ²	Adj. <i>R</i> ²	<i>SD</i> <i>residual</i>
CWIT-1	Intercept	9.863	[9.474, 10.253]	0.20	49.71	< .001		.155	2.775
	Age	-0.049	[-0.059, -0.039]	0.01	-9.89	<.001	.088		
	Age ²	-0.001	[-0.002, <-0.00]	<0.01	-3.12	.002	.010		
	Female	0.825	[0.456, 1.193]	0.19	4.39	<.001	.019		
CWIT-2	Intercept	10.217	[9.919, 10.515]	0.15	67.27	< .001		.031	2.797
	Age	-0.019	[-0.028, -0.009]	0.01	-3.76	<.001	.014		
	Age ²	<-.001	[-0.002, <-0.001]	<0.01	-2.80	.005	.008		
CWIT-3	Intercept	10.182	[9.824, 10.541]	0.18	55.67	< .001		.291	2.546
	Age	-0.073	[-0.081, -0.064]	0.01	-15.97	<.001	.202		
	Age ²	-0.001	[-0.002, <-0.001]	<0.01	-3.89	<.001	.015		
	Edu	0.078	[0.022, 0.133]	0.03	2.75	.006	.008		
	Female	0.454	[0.116, 0.793]	0.18	2.64	.009	.007		
CWIT-4	Intercept	10.561	[10.284, 10.838]	0.14	74.84	< .001		.250	2.574
	Age	-0.063	[-0.072, -0.054]	0.01	-13.83	<.001	.160		
	Age ²	-0.002	[-0.002, <0.001]	<0.01	-5.01	<.001	.024		
	Edu	0.098	[0.042, 0.154]	0.03	3.43	<.001	.012		

Note. *b* = unstandardized beta coefficient; *s.e.* = standard error of the unstandardized beta coefficient; *SD residual* = standard deviation of the residuals; Sex was coded 0 = men, 1 = women; Age and Education were mean centered, thus Age = (age - 46.2); Education = (the number of years of education obtained - 15.5); CWIT scores were transformed to scaled scores (*M* = 10, *SD* = 3) where higher scaled scores indicate increased test performance (Table 3).

Table 5

Total errors on CWIT-3 and CWIT-4 in a subset of the normative sample ($n = 935$)

Cumulative percentages		
<u>Errors</u>	<u>CWIT-3</u>	<u>CWIT-4</u>
0	100	100
1	51.4	55.4
2	23.2	26.5
3	11.1	12.9
4	5.2	6.3
5	2.8	3.2
6	1.2	1.5
7	0.5	1.1
8	0.4	0.6
9	0.2	
10		
11	0.1	0.2

Note. Cumulative percentages show proportion of the normative sample that attained k number of errors (or more).

Table 6

Results from multiple regression analysis on T-scores calculated with the original D-KEFS norms in the normative group (n = 1011)

Variable	Predictor	Original D-KEFS norms			Adj. R^2
		b	p	Partial R^2	
CWIT-1	Intercept	47.913	<.001		.023
	Age	0.037	<.004	.008	
	Education	-0.073	.380	.001	
	Sex	2.308	<.001	.020	
CWIT-2	Intercept	51.767	<.001		.030
	Age	0.069	<.001	.032	
	Education	0.102	.195	.002	
	Sex	0.442	.362	.001	
CWIT-3	Intercept	53.465	<.001		.016
	Age	0.039	.005	.008	
	Education	0.263	.004	.008	
	Sex	1.418	.012	.006	
CWIT-4	Intercept	52.710	<.001		.023
	Age	0.047	<.001	.011	
	Education	0.382	<.001	.017	
	Sex	0.762	.177	.002	

Note. b = unstandardized regression coefficient; p = p -value; partial R^2 = explained variance of predictor variable; Adj. R^2 = explained variance of combined predictor variables; significant coefficients ($p > .05$) indicate mal-adjustment in the norms; Age and education was mean centered

Table 7

Paired sample t-tests between T-scores computed using the Norwegian norms and original age-adjusted norms from D-KEFS

	M (SD)	t	p	M diff	95% CI M diff	Cohen's d
D-KEFS norms CWIT-1	49.5 (7.7)	-5.085	<.001	-0.55	[-0.76, -0.34]	-0.16
D-KEFS norms CWIT-2	52.1 (7.3)	16.055	<.001	2.06	[1.83, 2.30]	0.54
D-KEFS norms CWIT-3	54.4 (8.4)	38.267	<.001	4.41	[4.19, 4.64]	1.20
D-KEFS norms CWIT-4	53.2 (8.5)	25.436	<.001	3.22	[2.97, 3.47]	0.80

Note. $Df = 1010$; T-scores computed using Delis et al. (2001) norms were always paired with Norwegian norms that had a mean of 50 ($SD = 10$).

Table 8

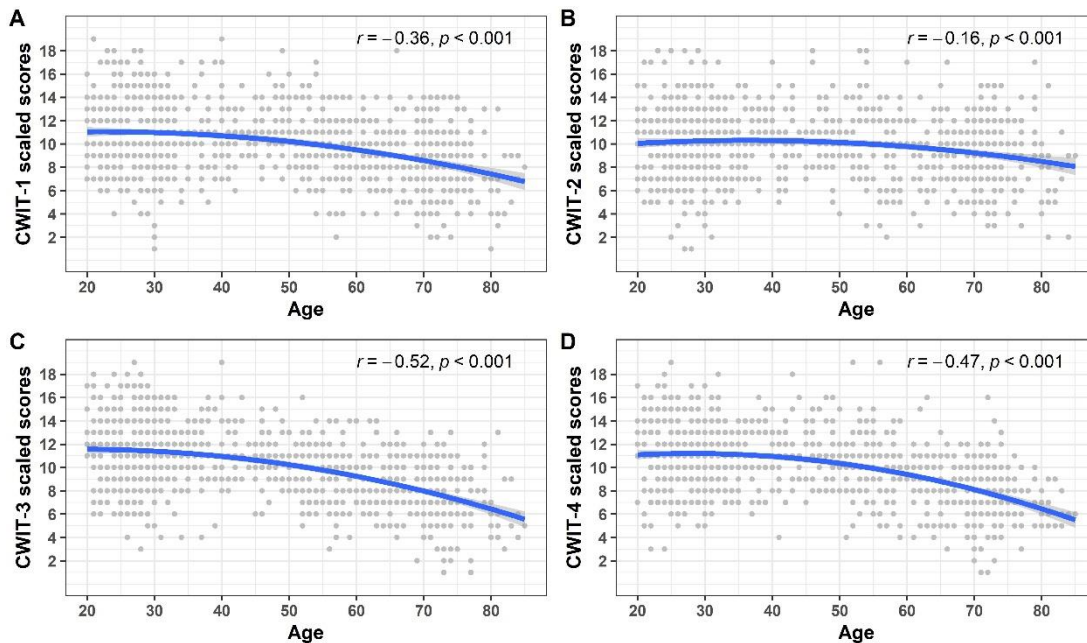
Intra-class correlations between baseline and follow-up on D-KEFS CWIT subtests based on a sub-set of the normative sample (n = 335).

Measure	ICC	95% CI [LL, UL]
CWIT-1 D-KEFS norms	.69	[.63, .74]
CWIT-1 Norwegian norms	.68	[.62, .74]
CWIT-2 D-KEFS norms	.62	[.55, .68]
CWIT-2 Norwegian norms	.68	[.61, .73]
CWIT-3 D-KEFS norms	.73	[.67, .78]
CWIT-3 Norwegian norms	.76	[.71, .80]
CWIT-4 D-KEFS norms	.66	[.59, .71]
CWIT-4 Norwegian norms	.70	[.64, .75]

Note. ICC = intraclass correlation coefficient; ICC (2,1) type.

Figure A.1

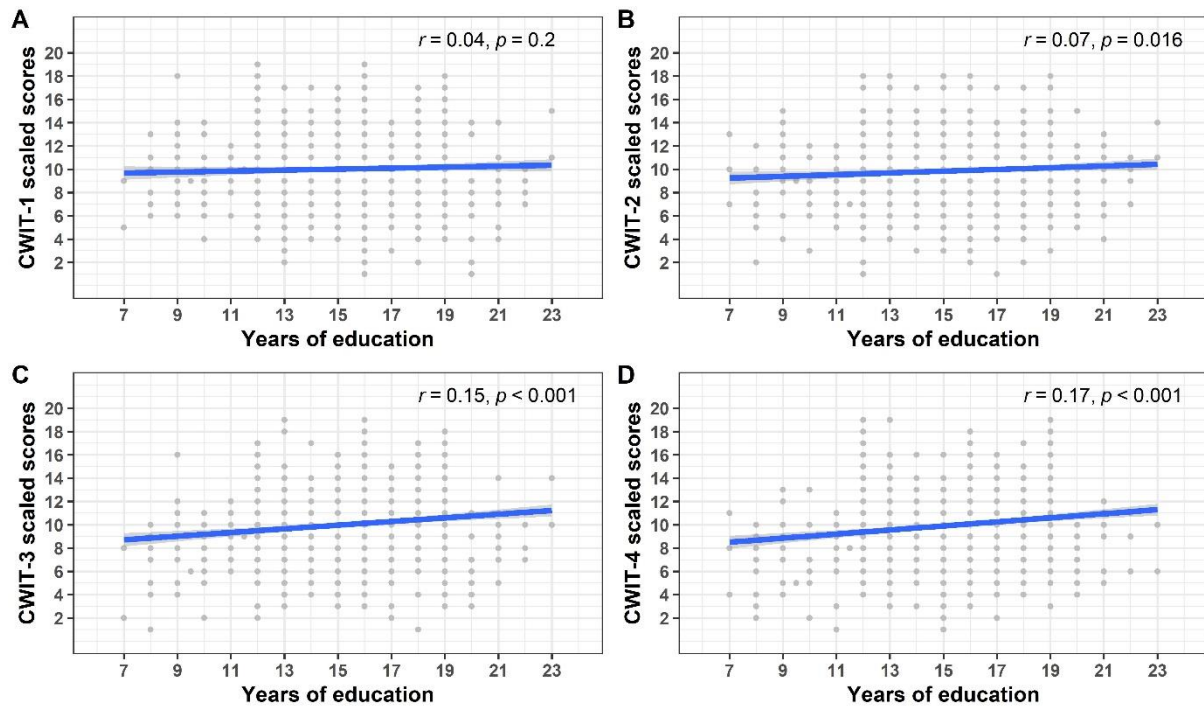
Scatterplots of age² and CWIT 1-4 scaled scores in the normative sample (n = 1011).



Note. r = Pearson correlation coefficient; p = p -value; blue line shows the regression line for the squared effect of age (age²); CWIT raw scores were transformed to scaled scores ($M = 10$, $SD = 3$) according to ranges described in Table 3.

Figure A.2

Scatterplots of years of education and CWIT 1-4 scaled scores in the normative sample ($n = 1011$).



Note. r = Pearson correlation coefficient; p = p -value; blue line shows the regression line between years of education and CWIT scaled scores; CWIT raw scores were transformed to scaled scores ($M = 10$, $SD = 3$) according to ranges described in Table 3.

Table A.1

Equality of age coefficients on CWIT subtests

Coefficients to be contrasted	b diff	$s.e.$	Z	p	b diff 99% CI [LL, UL]
CWIT-1 Age ($b = -0.049$) CWIT-2 Age ($b = -0.019$)	-0.031	0.004	-7.214	<.001	[-0.042, -0.020]
CWIT-1 Age ($b = -0.049$) CWIT-3 Age ($b = -0.073$)	0.024	0.004	5.608	<.001	[0.013, 0.035]
CWIT-1 Age ($b = -0.049$) CWIT-4 Age ($b = -0.063$)	0.014	0.005	2.698	.007	[0.001, 0.027]
CWIT-2 Age ($b = -0.019$) CWIT-3 Age ($b = -0.073$)	0.055	0.005	11.45	<.001	[0.042, 0.067]
CWIT-2 Age ($b = -0.019$) CWIT-4 Age ($b = -0.063$)	0.045	0.005	8.926	<.001	[0.032, 0.057]

Note. b = unstandardized beta coefficient; $s.e.$ = Standard error of b diff.

