Predicting cognitive development in people with multiple sclerosis using baseline processing speed

A six year prospective longitudinal study

Matthias Rasmussen Fuglestad & Jacob Ring

Neuropsychology

Submitted as Master's Thesis at the Department of Psychology

Faculty of Social Sciences

UNIVERSITY OF OSLO

Fall semester of 2022

Summary

Authors: Matthias Rasmussen Fuglestad & Jacob Ring
Title: Predicting cognitive development in people with multiple sclerosis using baseline processing speed. A six year prospective longitudinal study.
Supervisors: Nils Inge Landrø & Einar August Høgestøl

Background: Multiple sclerosis (MS) is a degenerative neurological disease that affects the central nervous system. Studies estimate a prevalence rate of cognitive impairment in between 43% and 70% of the MS population over the course of the disease. Processing speed (PS) is considered the most or one of the most affected cognitive domains in MS. However, there is a need for more prospective longitudinal studies of cognitive development in people with MS. Therefore, the goals of this study were to 1) Investigate if baseline PS could predict cognitive decline in people with MS and 2) Investigate if baseline PS could predict cognitive dysfunction in MS patients.
Methods: This study was done as a part of the Oslo Longitudinal MS cohort project. One of the authors contributed with data collection. A group of 76 people with newly diagnosed MS using a longitidunal prospective design were tested with a battery consisting of the Symbol digit modalities test, the Paced auditory serial addition test, the Color-word interference test, the Controlled oral word association test, the Brief visual memory test, the California verbal learning test second edition and some subtests of the Wechscler abbreviated scale of intelligence. We took two statistical approaches, a linear regression-based model and a linear mixed-effects based model. We chose to use baseline PS as our predictor with variables like other cognitive domains, BDI, FSS, IQ, EDSS & cognitive reserve included in our models when attempting to predict later cognitive outcome.
Results: Over the first two follow-ups, there was a trend of increased performance on neuropsychological tests. On the third follow-up, there was a drop in performance relative to the previous follow-up. We found a significant association with a small to medium effect size (*d* =-0.519) between baseline PS and cognitive dysfunction. We could not find this when controlling for some known confounding variables. Further, we investigated whether a dysfunction in PS at baseline could predict later cognitive dysfunction. There was a significant and moderate effect size (*d* = 0.738) for this relationship. We could not find a statistically significant association between baseline PS and later cognitive decline.

Discussion & conclusion: Baseline PS could not predict cognitive decline. However, baseline PS was significantly related later cognitive dysfunction. We interpret this finding as that performance on cognitive tests are relatively stable over time. EF emerges as a better predictor compared to PS in relation to later cognitive outcome. The pattern of cognitive test performance deviates from earlier MS research litterature. We argue that this is due to our group being an example of the modern MS patient, as practice effects alone cannot fully explain the trend. Additionally, the participant group is highly educated. Further, the defining characteristic of those fulfilling the criteria for decline was an increase in BDI score from T1 to T4 and lower EF T-score at T1. We could not find additional baseline data that can be used for prediction of later cognitive outcome.

Table of contents

Tables

Figures

Abbreviations

(Norwegian in *italics*)

BDI	Beck depression inventory
BICAMS	Brief international cognitive assessment for multiple sclerosis
BRB-N	Brief repeatable battery of neuropsychological tests
BVMT-R	Brief visuospatial memory test-revised
CNS	Central nervous system
COWAT	Controlled oral word association test
CSF	Cerebrospinal fluid
CVLT-II	California verbal learning test second edition
D-KEFS	Delis-Kaplan executive function system
EDSS	Expanded disability status scale
EF	Executive function
EM	Expectation-maximization
FDA	Food and drug administration
FSS	Fatigue severity scale
GCF	Global cognitive function
IQ	Intelligence quotient
LME	Linear mixed effects
MACFIMS	Minimum assessment of cognitive function in multiple sclerosis
MRI	Magnetic resonance imaging
MS	Multiple sclerosis
PASAT	Paced auditory serial addition test
PPMS	Primary progressive multiple sclerosis
PS	Processing speed
RRMS	Relapsing remitting multiple sclerosis
SDMT	Symbol digits modalities test
SPMS	Secondary progressive multiple sclerosis
SSB	*Statistisk sentralbyrå*
T1, T2, T3 and T4	Time of assessment 1, 2, 3 and 4
TP	Time point

WAIS-IV	Wechsler adult intelligence scale fourth edition
WASI	Wechsler abbreviated scale of intelligence
WISC-IV	Wechsler intelligence scale for children fourth edition

1 Introduction

1.1 An introduction to multiple sclerosis

Multiple sclerosis (MS) is a degenerative neurological disease that affects the central nervous system (CNS). Hallmarks of the disease include damage to the myelin sheath around the axon of neurons and axonal degeneration (Compston & Coles, 2008). The etiology of the disease is not fully known. However, factors that are related to the development of the disease include both genetic and environmental factors such as: low vitamin D, low exposure to UVB sunlight, toxins, smoking, diet, positive biomarkers of the Epstein-Barr virus, and distance to the equator in early childhood (Dobson & Giovannoni, 2019; Doshi & Chataway, 2017). However, systematic differences in the prevalence of MS in the southern and northern parts of Norway have not been found (Berg-Hansen, Moen, Harbo & Celius, 2014). A recent 2022 study confirms that the Epstein-Barr virus is a mandatory infection to develop MS and the main risk factor of the disease (Bjornevik et al., 2022). The age of onset is typically between 20 and 40 years, also women are diagnosed more often than men (Dobson & Giovannoni, 2019).

The disease may affect all different parts of the brain and spine and manifest with differently clinically observed patterns between individuals. Common symptoms of MS include symptoms related to mobility, vision, fatigue, cognition, bladder and bowel function, sensory function, spasticity, pain, depression, tremor, and coordination (Kister et al., 2013; Compston & Coles, 2008). MS may manifest in different clinical courses categorized into different subtypes. The participants in this study were classified into one of three subtypes following the 2010-criteria (Klineova & Lublin, 2018): "relapse remitting multiple sclerosis" (RRMS), "primary progressive multiple sclerosis" (PPMS), and "secondary progressive multiple sclerosis" (SPMS). RRMS is the most common subtype of MS and accounts for 80-85% of initial diagnoses (Milo & Miller, 2014) and is characterized by recurring attacks that include either new neurologic symptoms or worsening of existing neurologic symptoms. RRMS attacks are followed by a partial or complete recovery period between relapses. PPMS is characterized by a progressive increase in symptoms with occasional pauses in disease progression and some periods with minor improvement (Milo & Miller, 2014). SPMS starts as

RRMS with attacks and recovery periods followed by an additional onset of a progressive decline in neurologic function, either with or without attacks, with occasional pauses in disease progression and minor improvements (Milo & Miller, 2014).

There is no single diagnostic factor; however, diagnosis is primarily based on objective clinical findings of damage to the CNS at two different time points and primarily in two different locations (Milo & Miller, 2014). MRI and CSF analysis can be used as supplementary examinations. Anamnesis is also a part of the diagnostic process. The clinician should also perform assessments to evaluate differential diagnoses (Helsedirektoratet, 2019; Milo & Miller, 2014).

Over the last 25 years, there have been considerable advancements in the treatment of MS. Many new and effective drugs like natalizumab and fingolimod have been approved, significantly reducing the impact of the disease on many patients (Tintore, Vidal-Jordana & Sastre-Garriga, 2019). In addition, in 2018, the drug ocrelizumab was approved, which was the first medicinal drug approved that could benefit MS patients with the primary progressive subtype (Tintore et al., 2019). Some longitudinal studies have found increased survival rates in people with MS over the last decades (Lunde, Assmus, Myhr, Bø & Grytten, 2017; Simonsen et al., 2021). This increased survival rate is presumably due to the advancements in medicinal treatment; however, studies on MS survival rates over time are conflicting. Some studies show no reduced mortality if one accounts for the increase in survival rate in the general population (Kingwell et al., 2012). However, this modern breakthrough in treating MS is changing the long-term outcome for people with MS, mainly by slowing disease progression (Dobson & Giovannoni, 2019).

1.2 MS and cognition

Studies estimate a prevalence rate of cognitive impairment in between 43% and 70% of the MS population over the course of the disease (Chiaravalloti & Deluca, 2008). The cognitive symptoms may greatly vary between individuals (Patti, 2009). Different review articles highlight, to a certain degree, different cognitive domains as the most prominent domains affected by the disease. However, all articles mention processing speed (PS) as the most or one of the most affected cognitive domains in MS (Chiaravalloti & Deluca, 2008; Patti, 2009; Sumowski et al., 2018). Other cognitive domains mentioned as most commonly affected are episodic memory, long-term memory, learning, attention, executive function (EF), verbal fluency, and visuospatial analysis (Patti, 2009; Chiaravalloti & Deluca, 2008; Sumowski et al., 2018). The fact that a large part of the cognition is highlighted as areas commonly affected reflects the diversity of cognitive symptoms in MS. However, deficits in language and the development of dementia are rare in the MS population (Chiaravalloti & Deluca, 2008; Patti, 2009). Other uncommon cognitive symptoms are impairment in essential verbal skills, e.g., word comprehension and naming, and "simple" attention like repeating digits (Chiaravalloti & Deluca, 2008). General intelligence remains typically intact, but some studies find slightly reduced general intelligence (Chiaravalloti & Deluca, 2008).

It would be reasonable to assume that there will be additional cognitive decline as the disease progresses. However, studies investigating this have given conflicting results (Patti, 2009). This is presumably due to a lack of comprehensive and comparable studies investigating cognition in MS, as well as difficulties with interpreting available data due to heterogeneity in cognitive symptoms, duration of cognitive symptoms, disease course, and treatment, as well as a lack of follow-up studies (Patti, 2009). Sumowksi et al. (2018) state a need for studies with a prospective longitudinal design to further investigate the disease course-related cognitive decline rather than studies with a cross-sectional design. Despite this lack of conclusive evidence, it is still believed that cognitive impairment increases with the progression of the disease (Patti, 2009). A recent longitudinal study by Katsari et al. (2020) confirms this. They assessed cognitive function in people with MS over ten years and found that the overall proportion of participants with cognitive impairment increased by 10%.

Studies have also explored the relationship between physical disability, usually measured with the Expanded Disability Status Scale (EDSS), and cognition in people with MS. Some studies find no correlation, and some observe a low correlation between physical disability and cognitive dysfunction (Chiaravalloti & Deluca, 2008; Sumowski et al., 2018; Patti, 2009; Katsari et al., 2020). Studies exploring correlations in cognitive dysfunction across different cognitive domains in people with MS find no correlation or low correlation (Sumowski et al., 2018). Longitudinal brain imaging studies investigating associations between brain atrophy and cognitive decline in people with MS find moderate to strong correlations between an increase in atrophy and a change in cognitive function (Chiaravalloti & Deluca, 2008; Patti, 2009). A study investigating the relationship between PS, fatigue, depression, and cognition in people with MS (Diamond, Johnson, Kaufman & Graves, 2008) found that PS was significantly and moderately correlated with depression, fatigue, verbal fluency, and verbal memory. To summarize, studies indicate that cognitive function in MS is significantly correlated with brain atrophy, fatigue, and depression. There is no or a weak correlation between cognition and physical disability, and PS is moderately correlated with depression, fatigue, verbal fluency, and verbal memory.

1.3 MS and processing speed

As mentioned, PS is one of the most commonly affected functions in people with MS (Chiaravalotti & Deluca, 2008). Reduced PS is a development that resonates well with the mechanism of demyelination and cortical attacks. This mechanism has inspired the idea that reduced PS is one of the earliest signs of cognitive decline we can observe in people with MS. Stephen Rao and colleagues were the first to investigate the differences in PS in people with MS compared to controls (Rao, Leo, Haughton, Aubin-Faubert & Bernadin, 1989). The idea was that language and general intellectual ability appear to be preserved, so participants' cognitive changes should be related to cognitive functions such as PS, memory, or abstract reasoning. Rao and colleagues chose the former cognitive domain and found that the participants with MS were 47 % slower than controls on reaction time measures.

PS in neuropsychology is about the efficiency of our cognitive processes. Assessing efficiency is typically done by looking at functions like reaction time, completion time of tasks with simple instructions or the number of correct responses within a set time of a simple task. The Symbol Digit Modalities Test (SDMT) and Paced Auditory Serial Addition Test (PASAT) are examples of the latter.

Everyday life is a complex sequence of many different cognitive tasks to every one of us. The more demanding tasks require us to mobilize more cognitive resources, while less challenging tasks take us little to no conscious effort. So, what happens when the low-demand tasks require more effort than before? Salthouse proposed that the decline in PS function related to aging is a significant component of age-related differences in cognition (Salthouse, 1996). Following up on this, a longitudinal study on the effect of PS on fluid intelligence found a pearson correlation of 0.53 between the change in fluid intelligence and the change in processing speed over four years (Zimprich & Martin, 2002). The study also replicated Salthouse and Verhaegens's finding of a correlation of 0.52 between PS and fluid intelligence (Verhaeghen & Salthouse, 1997). These findings suggest that changes in multiple cognitive domains across the lifespan are related to PS differences.

PS has been pointed to as a primary deficit in MS (Demaree, DeLuca, Gaudino & Diamond, 1999). In a 12-year longitudinal study using survival curves, the first two tests where people with MS failed to reach the 5th percentile of the normal population were tests mainly measuring PS (Van Schependom et al., 2015). PS has also been shown to influence employment status, an important aspect of MS patients lives (Strober et al., 2012). The prevalence of PS deficits also seems to appear across different subtypes of MS, with some debate about which subtypes are more or less affected (Costa, Genova, DeLuca & Chiaravalloti, 2017).

When assessing people with suspected or recently diagnosed MS, neuropsychological testing is standard, and many different test batteries have been proposed. At the start of the millennium, a group of experts came together to create a minimal neuropsychological assessment of MS. Their finalized efforts resulted in the Minimum Assessment of Cognitive Function in Multiple Sclerosis (MACFIMS) battery assessing five areas of cognitive function: PS, learning/memory, EF, visuospatial processing, and word retrieval (Benedict et al., 2006). PS was allotted two of the seven tests in the battery, a testament to its importance.

Given the view of MS as a progressive neurological disorder, and the finding that the first two tests most often failed by MS patients were tests mainly measuring PS (Van Schependom et al., 2015), we wanted to look into whether PS could predict later cognitive function. A previous study found that better performance on SDMT at baseline decreased the probability of scoring 4.0 or higher on EDSS at 10-year follow-up (Moccia et al., 2016). This was also the only test still significant after controlling for age and baseline EDSS. A later study by Hechenberger et al. also found that SDMT at baseline could predict EDSS results seven years later (Hechenberger et al., 2022). This indicates there is some support for using baseline SDMT as a measure of PS to predict later disability in people with MS. A study by Bergendal, Fredrikson & Almkvist (2007) found that early cognitive impairment predicts further cognitive decline. They also found that PS was an especially strong predictor of later cognitive decline.

1.4 Research questions and hypotheses

The main aim is to investigate the relationship between baseline PS in people with MS and later cognitive outcomes. A potential finding may contribute to an early indication of later cognitive progression, increasing the clinicians' prerequisites for making a more accurate prediction of later cognitive function in people with MS.

PS presents as a promising cognitive domain concerning the prediction of later cognitive function. This is mainly due to two factors. The first is that the demyelination disease mechanism previously explained resonates well with our hypothesis that baseline scores on neuropsychological tests that measure PS may indicate a more severe progression of cognitive decline over time. The second factor is that in our literature review, the studies investigating cognition in MS usually find that the most robust associations are related to PS (see Bergendal et al., 2007; Benedict, Morrow, Guttman, Cookfair & Schretlen, 2010).

Cognitive dysfunction is often defined in a clinical setting with reference to standardized scores below a cut-off point. Cognitive decline refers to a reduction in the participants' scores on neuropsychological tests relative to the baseline measurement. This thesis will investigate the following two research questions:

1. Can baseline processing speed predict later cognitive decline in people with MS?
2. Can baseline processing speed predict later cognitive dysfunction in people with MS?

By exploring both cognitive dysfunction and decline, we can additionally investigate changes in cognition in participants who have a reduction in cognitive function relative to baseline measures in our follow-ups but do not meet the criterion for cognitive dysfunction as it is operationalized in this thesis.

We use two different neuropsychological tests to assess PS in this study, the SDMT and the PASAT 3" (the three-second interval version). We will mainly look at their combined predictive properties in relation to our outcome variables. However, we will also look at them separately to investigate if they have different predictive properties. We will also examine the relationship between baseline measurements of verbal fluency, verbal memory, visual memory, EF, and later cognitive outcome, but we will mainly focus on PS. In addition, to attempt to predict cognitive decline and cognitive dysfunction, we will also investigate a global score for cognitive function to get more general information about baseline cognitive function and later cognitive function. We have designed this study in an effort to make it clinically relevant.

2 Method

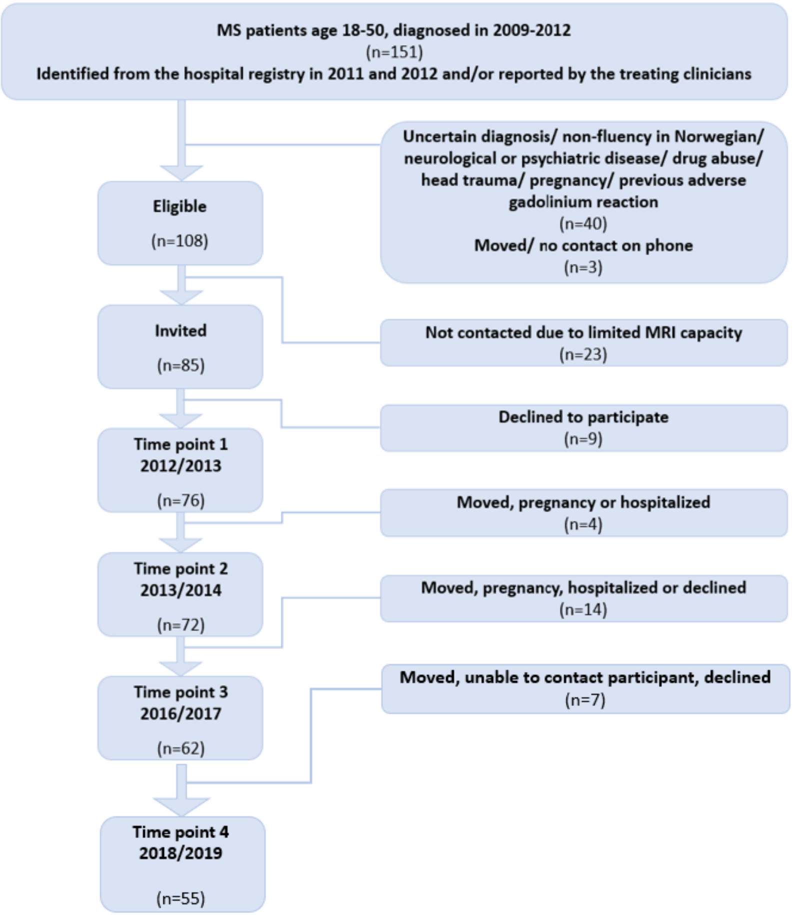
2.1 Participant selection process

Patients affiliated with the Department of Neurology at Ullevål hospital, a part of the Oslo University Hospital, were assessed as candidates for a prospective longitudinal study (Høgestøl, 2020). The patients assessed were between 18 and 50 years of age and had RRMS diagnosed between 2009 and 2012. A total of 151 patients were assessed for inclusion in this study (see Figure 1). In the end, 43 patients did not get invited to participate due to fulfillment of at least one exclusion criterion. The exclusion criteria were difficulties with speaking the Norwegian language, earlier adverse reaction to MRI contrast agents, uncertain diagnosis, substance abuse, other neurologic or psychiatric diseases, pregnancy, or earlier traumatic brain injuries (Høgestøl, 2020). Of the 108 remaining candidates that passed the inclusion criteria, 85 were selected due to limitations in the MRI scanner capacity. Of the 85 candidates invited to participate in the study, nine declined for various reasons listed in Figure 1.

Over the three follow-ups (T2, T3, and T4), some participants could not continue participating in this study. The five main reasons stated by the participants for this were: pregnancy or young children at home, the participant moving to another city or country, the participant being hospitalized, lack of time, and the participant didn’t want to continue to participate in the study (Høgestøl, 2020). At the baseline measurement in 2012/2013, there were 76 participants. At the first follow-up in 2013/2014, there were 72 participants. At the second follow-up in 2016/2017, there were 62 participants. Finally, at the third follow-up in 2018/2019, there were 56 participants. The reduction of participants over time and the selection of participants due to limitations in MRI scanner capacity may have contributed to a selection bias in this dataset.

Figure 1

The selection process and decline of study participants (Høgestøl, 2020)


MS patients age 18-50, diagnosed in 2009-2012
(n=151)
Identified from the hospital registry in 2011 and 2012 and/or reported by the treating clinicians
Uncertain diagnosis/ non-fluency in Norwegian/ neurological or psychiatric disease/ drug abuse/ head trauma/ pregnancy/ previous adverse gadolinium reaction
(n=40)
Moved/ no contact on phone
(n=3)
Eligible
(n=108)
Not contacted due to limited MRI capacity
(n=23)
Invited
(n=85)
Declined to participate
(n=9)
Time point 1
2012/2013
(n=76)
Moved, pregnancy or hospitalized
(n=4)
Time point 2
2013/2014
(n=72)
Moved, pregnancy, hospitalized or declined
(n=14)
Time point 3
2016/2017
(n=62)
Moved, unable to contact participant, declined
(n=7)
Time point 4
2018/2019
(n=55)

2.2 Neuropsychological tests and psychometrics

The tests used in this study were administered to the participant group at zero, one, four, and six years after being diagnosed with MS. Administering these tests give us a way to quantify each person on the cognitive domains and use this data to investigate our research questions. In order for this to happen, the tests have to measure and quantify the domains we want to investigate and have adequate psychometric properties. From the six neuropsychological tests used in this study, we usually got 14 different cognitive variables per participant at each of the four timepoints (TPs) participants got tested. We will now give a brief overview of the tests we've used and the cognitive domains they measure.

The California Verbal Learning Test second edition (CVLT-II) is used to measure learning and memory of verbal information and the use of strategies and processes that promote the aforementioned (Delis, Kramer, Kaplan & Ober, 1987). The learning trial consists of the examiner reading a list five times. After every read the participant tries to remember as many words as possible from the list and orally repeats them to the examiner. The total amount of correct responses over the five tries can be converted to a T-score for learning. For short-term memory, a second list is administered once in between, and then the participant tries to remember as many words as possible from the first list. The total correct answers in this condition can be converted to a T-score for short-delay recall. The long-delay recall condition was not administered in this study. Test-retest reliability for trials one to five and short-delay free recall yielded a reliability coefficient of 0.80 in the standard version (Woods, Delis, Scott, Kramer & Holdnack, 2006). This matches the manuals 0.82 quite well. The validity was compared to the original CVLT, and the results indicate that they measure the same construct. The construct validity has been demonstrated in over 200 studies, of which a summary exists in the manual.

Parts of the Wechsler Abbreviated Scale of Intelligence (WASI) were administered. Participants completed two subtests: matrix reasoning and vocabulary. This allows us to calculate an approximate intelligence quotient (IQ). The test has excellent psychometric properties with correlation coefficients ranging from 0.87-0.97 for internal consistency, 0.83-0.94 for test-retest, and 0.98-0.99 for interrater reliability. The concurrent validity with WASI, Wechsler Intelligence Scale for Children fourth edition (WISC-IV) and Wechsler Adult Intelligence Scale fourth edition (WAIS-IV) was measured between 0.71 to 0.92 (McCrimmon & Smith, 2013). This validity tells us that using the WASI will have a high amount of agreement to the full scale WISC and WAIS tests.

SDMT is a test where participants are given a paper with a symbol to number table on it. SDMT has a both written and oral version. The oral version is usually preferred for people with MS to avoid psychomotor interference as people with MS often have a reduction in motor function (Chen, Chiaravallotti, Genov & Costa, 2020), however the written version was used in this study. Participants are instructed that each symbol has a matching number and to fill in the correct number for the symbol in the tables below. They get 90 seconds to fill in as many as possible. The number of correct symbol to number answers can be converted to a T-score for PS. The manual of the SDMT claims a test-retest correlation of 0.80 in normal adults, which is considered good.

To get a second measure of PS the three-second version of the PASAT was administered. In this task, participants listen to a tape that reads a number every three seconds. The subjects are asked to continuously add the last read number to the number that immediately precedes it and orally present the answers. The total number of correct responses can be converted to a T-score for PS. The standard version of PASAT has good psychometric properties with internal consistency correlations ranging from 0.76 to 0.95, 0.96 for split-half reliability, and test-retest reliability correlations from 0.90 to 0.97 (Tombaugh, 2006). This means that the test has excellent reliability and good internal consistency.

The Brief Visual Memory Task Revised (BVMT-R) is a test measuring visuospatial memory. The administration features three learning trials where the participant gets to see a stimulus sheet for ten seconds per trial, before he/she is instructed to copy the figures onto a blank paper as accurately as possible. The delayed recall and recognition conditions were not administered in this study. According to PAR, the learning trials have interrater reliability of 0.96-0.97 and 0.97 for total recall (PAR, 2021). In a study validating the Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS) battery in Argentina the BVMT-R was found to have a test-retest correlation of 0.82 (Vanotti, Smerbeck, Benedict & Caceres, 2016). Correlations with other similar tests show that the test correlates more with the recall trials and tests of learning and memory. It had a low correlation to the Controlled Oral Word Association Task (COWAT) giving it discriminant validity (Benedict, Schretlen, Groninger, Dobraski & Shpritz, 1996).

COWAT is a test used to measure verbal fluency. There are three trials where the participant is given a letter and is asked to name as many words as possible starting with that letter in 60 seconds. The total correct words can be converted to a T-score for phonetic fluency. This condition is commonly referred to as the FAS trial, due to the letters used. After this a different subtest is administered, the participant is given a category and is asked to name as many words in that category as possible in 60 seconds. In a study testing reliability and validity of a shorter and a longer version the original FAS was found to have a test-retest correlation of 0.82 (Harrison, Buxton, Husain & Wise, 2000). This is similar to the 0.80 of the verbal fluency test in The Delis-Kaplan Executive Function System (D-KEFS) battery (Homack, Lee & Riccio, 2005).

D-KEFS Color Word Interference Test, hereafter mostly referred to as “Stroop” is mainly used to measure EF, but more specifically verbal inhibition and mental flexibility. It is a test comprised of four trials: 1) say the name of the color of five lines of squares in different colors, 2) read five lines of words spelling out different colors where the words are written in black with a white background, 3) name the ink color of words spelling out a different color over five lines, and 4) almost the same as condition three, however some words are boxed in. The boxed in words should be read and the ink color should be ignored. In a study using a sample of 101 with similar composition to the standardization sample a test-retest correlation of 0.62-0.76 was found (Homack, Lee & Riccio, 2005).

2.3 Standardizing the neuropsychological test raw scores

We standardized the SDMT, BVMT-R, CVLT-II, and Stroop using American norms found in the tests' respective manuals. We used Portuguese norms (Sousa, Neves, Passos, Ferreira & Sa, 2018) for the PASAT 3". COWAT was standardized using Norwegian norms from an article by Egeland, Landro, Tjemsland & Walbaekken (2005).

We standardized the SDMT, PASAT, BVMT-R, CVLT-II, and Stroop by age group. The SDMT, PASAT, COWAT, and Stroop were standardized by years of education. CVLT-II, SDMT, and PASAT were standardized by gender. Age was calculated for each participant at each TP as we knew the date of birth and the date of neuropsychological testing. We measured education at T1 and T3; however, only four participants had finished at least one additional year of education at T3 compared to T1.

The output of the standardized scores was in different units: T-scores, Z-scores, scaled scores, and percentiles. For a more straightforward overview and more straightforward analyses, we converted all standardized non-T-scores to T-scores. We did this using Microsoft Excel and making new variables using different conversion formulas. We used the formula "50+(10*Z-score)" for conversion from Z-score to T-score. For conversion from scaled score to T-score, we used the formula "50+((ScaledScore-10)*3)". We did not convert the percentiles as the variables standardized with percentiles in our analyses.

2.4 Operationalization of cognitive variables

For each measurement TP, we made a variable called "processing speed," which we operationalized as the sum of the participants' SDMT and PASAT T-scores divided by two, which gives us an average T-score of PS. At T2, PASAT was not administered. We solved this by using SDMT T-scores alone as the measure of PS at this TP. Some participants had only completed one of the two PS tests at each follow-up. This problem was solved by using only the completed test T-score as the score for PS at that TP for that participant.

We also made a variable for global cognitive function (GCF) at each TP. This was operationalized as the average T-score of the five cognitive domains PS, visual memory, verbal memory, phonetic fluency, and EF. By averaging the standardized scores of these cognitive domains, the results will be an equally weighted GCF variable. Operationalization of the PS variable is described above. Operationalization of visual memory is simply the T-score variable from the BVMT-R learning trials. We operationalized verbal memory as the average T-score of "learning" and "long-term memory" from CVLT-II. The operationalization of verbal fluency is simply the T-score from the COWAT F, A, and S phonetic fluency conditions. Finally, we operationalized EF as the average T-score of the Stroop completion time of conditions three and four.

We also made a dummy coded variable for "processing speed dysfunction" at each TP from our dataset. PS dysfunction was operationalized as PS T-score at or below 35, a common cut-off point for dysfunction (Lezak, Howieson, Bigler & Tranel, 2012, p. 172). We used the same dysfunction cut-off point for other cognitive domains as well. We also made a dummy coded variable for "cognitive dysfunction," which is meant to reflect a global cognitive dysfunction. Cognitive dysfunction was operationalized as a dysfunction in a minimum of two of our five cognitive domains for the participant at the TP. We argue that a dysfunction in one domain is too narrow and that dysfunction in three domains is too strict for this variable. We also made a dummy coded variable for decline in GCF from T1 to T4, where zero refers to roughly no change or positive change, and one refers to a decline in GCF from T1 to T4. Decline was operationalized as a minimum T-score change of -3 over this time interval.

We made a variable for cognitive and brain reserve that include years of education, WASI word comprehension T-scores, and normalized brain volume (the proportion of the volume inside the skull that is filled with brain tissue). MRI scans were used to obtain the latter data. Next, we weighted the variables roughly equally by dividing and multiplying each score with a common denominator for the average score of each variable, ending up with three variables with approximately the same median and arithmetic average values and standard deviations. We then summarized these three variables for each participant to make a cognitive and brain reserve variable. However, henceforth, we refer to this variable as cognitive reserve.

2.5 Ethical considerations

This study was approved by the Regional Committee for medical and health related research ethics - South East Norway. The participants were informed about the study both orally and in writing and signed informed consent forms before participation in this study. Inclusion in this study had the benefit of a more comprehensive follow-up of the participants' health status compared to what they would get without participation.

Cons include the time spent being examined in this study and some potentially uncomfortable examinations, such as MRI and straining neuropsychological tests. Participants had the opportunity to talk to healthcare professionals if they were anxious about the test results. The participants were offered breaks, and the TPs for examinations were flexible to ease the burden of participation. It is possible that some participants have experienced pressure to participate because the request for study participation was given by the hospital department responsible for their medical follow-up and treatment. The examination methods used in this study are generally considered safe and of low risk to participants.

2.6 Descriptive statistics

First, we made a table presenting descriptive information about our participant group. We divided our group by gender and looked at age, years of education, IQ, EDSS, and FSS scores to overview participant characteristics. In this table, we included the number of participants as well as range, mean, and standard deviation scores. Then we ran descriptive statistics for all the neuropsychological tests at all four TPs and presented this information in a table. We also made histograms to get a better overview of the score distribution for all neuropsychological tests and GCF T-scores at T1 and T4 and IQ at T1. Next, we ran descriptive statistics on our five cognitive domains and GCF scores at T1 and T4 and the mean T-score *changes* in our cognitive domains and GCF from T1 to T4 as well as a line diagram showing cognitive domain development over time. We also made histograms to see the distributions of the

T-score changes of our cognitive domains and GCF from T1 to T4. We made a bar graph that visually represents the number of participants with different combinations of PS intact (intact, meaning that the criterion for PS dysfunction is not met) and PS dysfunction at T1 and T4 to get an overview of the change tendencies in PS dysfunction over time in our participant group. Lastly we present a bar graph for different combinations of baseline PS dysfunction/intact and cognitive decline/no decline at T4.

2.7 Statistical analyses

Our linear regression analyses, which include most of our analyses, were performed in the statistics software IBM SPSS Statistics version 27 for Windows. Our linear mixed effects (LME) analyses were performed in the statistics software R version 4.1.2 for Windows and MAC, otherwise known as Bird-hippie. We have primarily been using the LME4 package to fit and analyse our models. The robustlmm package was also used to minimize the effects of outliers in our data. Some figures and tables were made using Microsoft Excel version 2201 for Windows.

2.7.1 Analyses to investigate the relationship between PS at T1 and global cognitive function at T4

First, we investigated the relationship between PS at T1 and GCF at T4. Regression and correlational analyses are well suited to explore this relationship. However, we choose mainly regression-based analyses as these are more naturally suited for prediction, which we are primarily looking into. Furthermore, our regression analyses were performed with an imputed dataset because regression analyses in SPSS use complete case analysis as default, which would result in the analyses ignoring 21 participants. To explore this relationship, we first performed a linear regression analysis in SPSS, where we inserted the GCF variable at T4 as the dependent variable and PS at T1 as the predictor. We then performed a multiple linear regression analysis to control for known confounding factors. In this analysis, we, in addition to PS at T1, inserted BDI, FSS, and EDSS scores at T1 and IQ, visual memory, verbal fluency, verbal memory, and EF at T1 as predictor variables. We then performed the same analysis but included cognitive reserve at T1 as a predictor to investigate the influence of higher levels of cognitive reserve on GCF at T4.

2.7.2 Analyses to investigate the relationship between PS at T1 and cognitive decline

The analyses mentioned above will give us a general idea of the relationship between baseline PS at T1 and later overall cognitive function. To directly answer our first research question regarding cognitive decline, we performed two more analyses: 1) a binary logistic regression analysis using PS at T1 as the predictor and a dummy coded variable for cognitive decline from T1 to T4 as the dependent variable. Binary logistic regression analyses are suited for this analysis as binary logistic regression can perform regression analyses using a numeric variable as a predictor and a dichotomous variable as the dependent variable. 2) We also investigated the relationship between PS dysfunction at T1 and the same cognitive decline variable using a "tests of independence chi-square" test for this analysis as a chi-square test is suited for investigating if there is a statistically significant relationship between two dichotomous variables. Finally, we included "Phi and Cramer's V" for an effect size analysis as a chi-square test alone does not give output on the strength of the relationship between variables in the analysis.

2.7.3 Analyses to investigate the relationship between PS at T1 and cognitive dysfunction at T4

We then explored the relationship between PS at T1 and cognitive dysfunction at T4, where the latter is a dummy coded variable. First, we performed this analysis by inserting PS at T1 as the predictor and cognitive dysfunction at T4 as the dependent variable using binary logistic regression analysis. We then performed the same analysis and inserted BDI, FSS, EDSS, IQ, visual memory, verbal fluency, verbal memory, and EF at T1 as additional predictor variables to control for these variables. Then we performed the latter analysis again but included cognitive reserve as a predictor. We also explored the relationship between PS dysfunction at T1 and cognitive dysfunction at T4. Both variables are dummy coded. We investigated this relationship using a “tests of independence chi-square” test including “Phi and Cramers’ V” for effect size data.

2.7.4 Additional repeated measurement longitudinal analyses

We also wanted to analyse the relationship between baseline PS and GCF at T4 using data from all four measurement points. This analysis takes full advantage of our repeated measurement longitudinal data and might find nuances in the relationship between these variables that we otherwise would miss. To do this, we performed an LME analysis. Here we use a robust version of LME modeling (the robustlmm package), meaning that in the scoring equations, residuals and predicted random effects are replaced with a bounded function (Koller, 2016). This limits the impact a single outlier can have on our total distribution. The analyses have also been done with a standard LME program: the Lmer from the LM4 package in R. In these models, we added the variable at our disposal, PS (in one analysis, PS was replaced with SDMT+PASAT as two separate predictors in the same model), EDSS, gender, EF, and EDSS. The reason we did both a standard and a robust analysis was to see how much controlling for outliers would impact the overall model. In all these analyses, the goal was to predict GCF at T4.

One of the perks of this model type is that it does not use complete case analysis as default and analyses data in a hierarchical structure. This approach means that the impact one missing data point has on an individual participants’ data is more negligible because participant data in this analysis builds on up to four observations. In addition, participants with one or more missing data points are, in essence, precision weighted, and extremes, therefore, shrink towards the mean (Brown, 2021). This weighting minimizes the bias caused by dropout. Because of this, we used the original data set with dropout for this analysis.

We used the lme.dscore function in the EMStools package for R statistic to get effect sizes for the models. This function instantly calculates Cohen’s *d* for all predictors used in a Lme4 or Lmer fitted model. To get the standard errors of the model, we used the summary function in R statistics to provide us with an estimated standard error and T-score of the predictors and intercept.

2.7.5 Supplementary analyses

To get effect size data between baseline PS and our cognitive outcome variables we performed a Pearson *r* correlation analysis and converted the Pearson *r* coefficient to Cohen's *d* by using the formula below (Ruscio, 2008).

$$d = 2 \cdot \frac{r}{\sqrt{1 - r^2}}$$

We performed separate analyses for SDMT and PASAT to investigate if there are any differences in their relationship with both later GCF and cognitive dysfunction. We did this investigation by running four separate linear regression analyses where SDMT and PASAT separately are inserted as predictors and GCF and cognitive dysfunction separately as the dependent variables.

We performed a linear regression analysis using cognitive reserve as predictor and GCF at T4 as the dependent variable to investigate if our cognitive reserve variable alone is statistically significantly related to our main cognitive outcome variables. We also performed a binary logistic analysis using cognitive reserve as the predictor and cognitive dysfunction at T4 as the dependent variable.

Most of our analyses are regression-based due to their predictive nature. A regression slope coefficient gives information about the strength of the relationship between the variables examined, and because our cognitive domain predictors are all standardized using the same standardized unit (T-scores), we argue that we can directly compare the different slope (unstandardized beta) coefficients from these domains to compare their relative relationship strength to our (also standardized by T-score) cognitive outcome variables. We do point out that the slope coefficient does not have correlational or effect size properties and is solely a value reflecting the slope of the regression line in a regression model.

2.7.6 Post-hoc analyses

We performed post-hoc analyses of the relationship between baseline EF and our cognitive outcome variables to further explore our findings. We performed a linear regression analysis using baseline EF as predictor and GCF at T4 as the outcome variable. We also performed a binary logistics regression analysis using baseline EF as predictor and cognitive decline as the outcome variable. Additionally, we performed a binary logistic regression analysis using baseline EF and cognitive dysfunction at T4 as the outcome variable. We followed the above-mentioned steps for calculating a Cohen's *d* for all these three analyses. Lastly, we performed LME-analyses to use our repeated measurement data and explore covariates affecting baseline EF and later GCF.

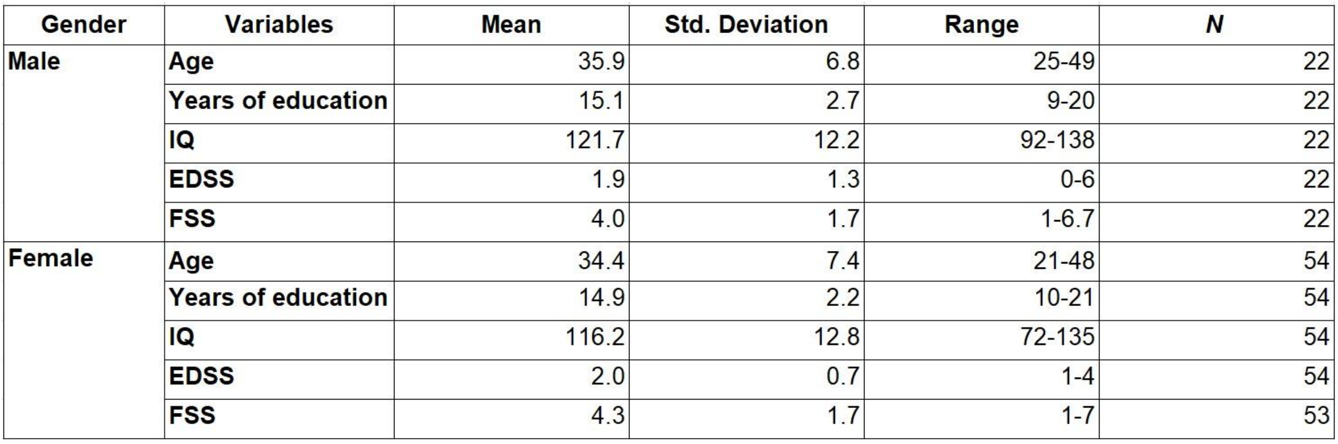
3 Results

3.1 Descriptive statistics

Descriptive information of our participant group at baseline is listed in Table 1. From this table, we see that the vast majority of the participant group are women (54 out of 76). There are no prominent gender differences in this table; however, it is worth mentioning that the minimum IQ of the men is 92 compared to 72 for women, and the average IQ for men is roughly 5.5 points higher compared to the average IQ of the women in this group. Average IQ and years of education are higher for both men and women in our participant group compared to the general population, with average IQ scores of roughly 122 for men and 116 for women, as well as average years of education of approximately 15 years for both men and women.

Table 1

Descriptive information of the participants at T1



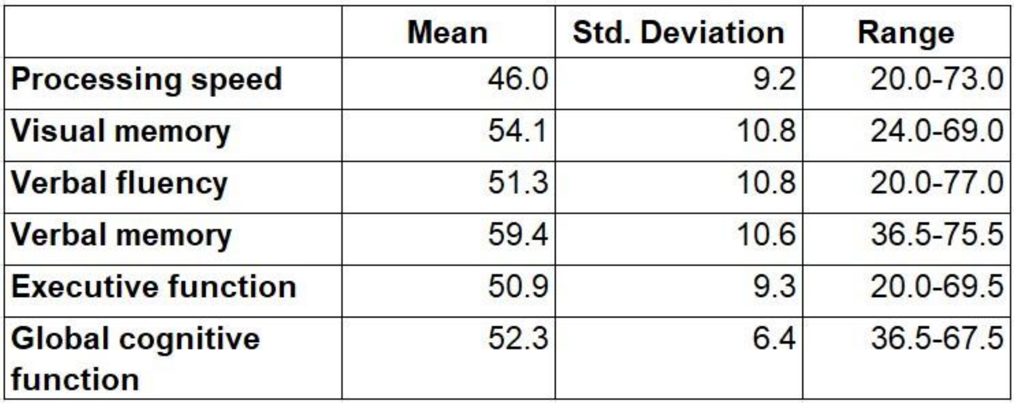
Gender	Variables	Mean	Std. Deviation	Range	*N*
Male	**Age**	35.9	6.8	25-49	22
	Years of education	15.1	2.7	9-20	22
	IQ	121.7	12.2	92-138	22
	EDSS	1.9	1.3	0-6	22
	FSS	4.0	1.7	1-6.7	22
Female	**Age**	34.4	7.4	21-48	54
	Years of education	14.9	2.2	10-21	54
	IQ	116.2	12.8	72-135	54
	EDSS	2.0	0.7	1-4	54
	FSS	4.3	1.7	1-7	53

See Figure 2 for group average T-scores of all cognitive domains and GCF at all TPs (descriptive statistics for all specific tests can be found in Table A1 in appendix A). A general pattern in this figure seems to be a gradual increase in test performance from T1 to T2 and from T2 to T3, followed by a reduction in test performance from T3 to T4, although not all tests at all follow-ups follow this pattern. Histograms of all neuropsychological tests and GCF T-score distribution at T1 and T4 can also be found in Appendix A (see Figures A1 to A21). Both Table A1 and Figures A1 to A21 are based upon the imputed dataset, as this dataset is used the most in our analyses. We also point out that descriptive statistics of the original and the imputed dataset are very similar.

Table 2 shows descriptive data of our five cognitive domains and GCF T-scores at T1. Table 3 shows the same for T4. Table 4 shows the mean *changes* in T-scores from T1 to T4 in the same domains. Figures 3 to 8 show histograms of the change scores in these domains from T1 to T4 to better understand the distribution. From Table 2, the most prominent score here is the verbal memory mean T-score of 59.4, which is roughly one standard deviation above the norm average of 50. We can also see that the mean visual memory T-score is above the norm average with a score of 54.1 and that the mean PS T-score is 46, which is below the norm average. Table 3 shows the same tendencies; however, in this table, the EF mean T-score is now above the norm average with a score of 54.3. In Table 4, we can see that the mean T-scores of EF and verbal memory from T1 to T4 have increased slightly. On the other hand, the average verbal fluency T-score has been reduced by 4.8 over the same period. PS, visual memory, and GCF have roughly the same average T-scores at T1 and T4. We also see a bigger spread in visual memory and verbal memory T-scores at T4 relative to T1.

Table 2

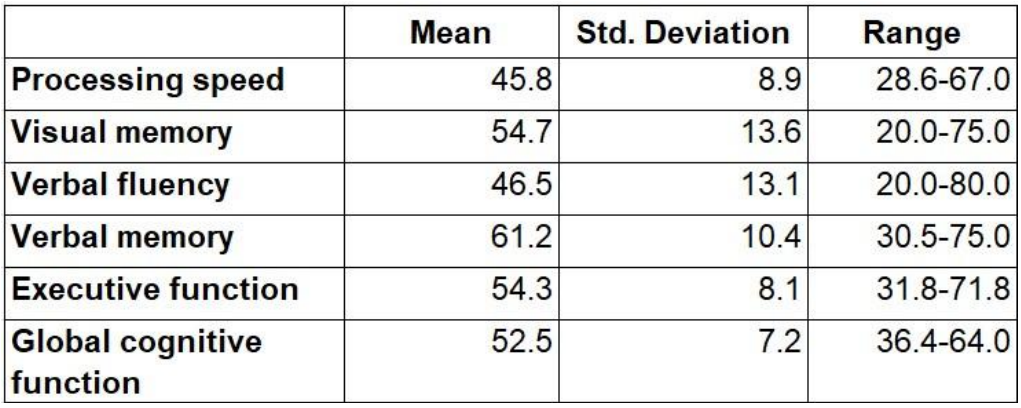
Descriptive statistics of cognitive domains and GCF at T1, given in T-scores.



	Mean	Std. Deviation	Range
Processing speed	46.0	9.2	20.0-73.0
Visual memory	54.1	10.8	24.0-69.0
Verbal fluency	51.3	10.8	20.0-77.0
Verbal memory	59.4	10.6	36.5-75.5
Executive function	50.9	9.3	20.0-69.5
Global cognitive function	52.3	6.4	36.5-67.5

Table 3

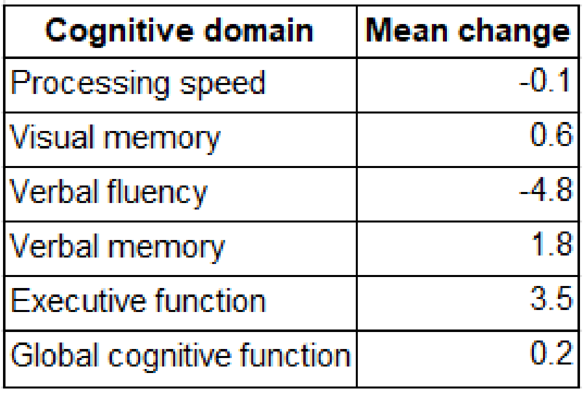
Descriptive statistics of cognitive domains and GCF at T4, given in T-scores.



	Mean	Std. Deviation	Range
Processing speed	45.8	8.9	28.6-67.0
Visual memory	54.7	13.6	20.0-75.0
Verbal fluency	46.5	13.1	20.0-80.0
Verbal memory	61.2	10.4	30.5-75.0
Executive function	54.3	8.1	31.8-71.8
Global cognitive function	52.5	7.2	36.4-64.0

Table 4

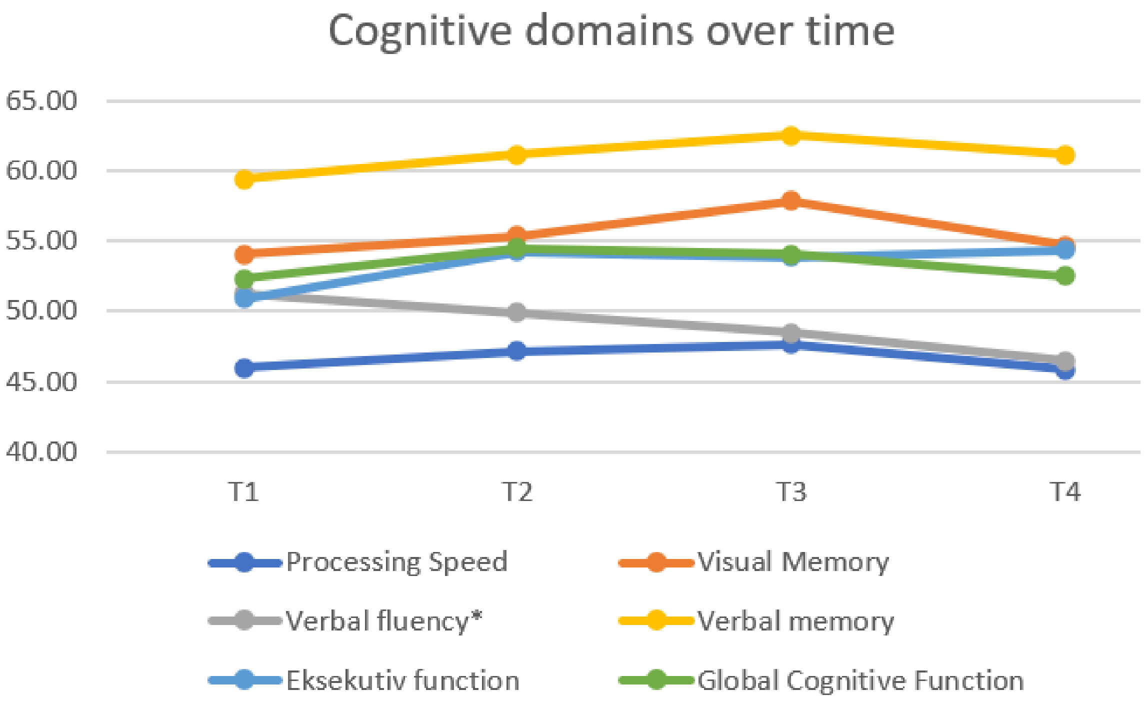
Descriptive statistics of cognitive domains and GCF T-score changes from T1 to T4 (positive scores refer to a T-score increase over the follow-up period, and negative scores refer to a decline)



Cognitive domain	Mean change
Processing speed	-0.1
Visual memory	0.6
Verbal fluency	-4.8
Verbal memory	1.8
Executive function	3.5
Global cognitive function	0.2

Figure 2

Line diagram showing the measured cognitive domains over the different timepoints. All scores are group averages of T-scores.



**Verbal fluency is missing at T2 and is for continuity here shown as the average of T1 and T3*

Figure 3

Histogram of PS T-score change from T1 to T4 (positive scores refer to a T-score increase over the follow-up period, and negative scores refer to a decline).

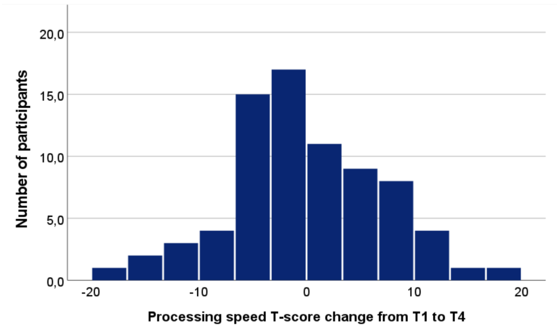


Figure 4

Histogram of EF T-score change from T1 to T4 (positive scores refer to a T-score increase over the follow-up period, and negative scores refer to a decline).

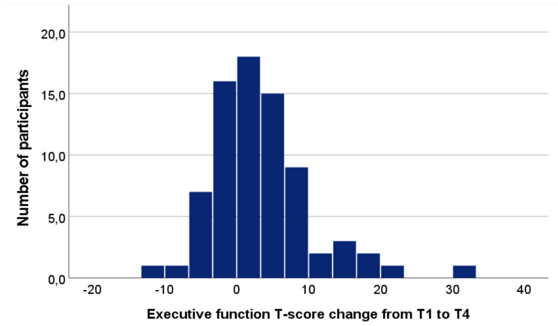


Figure 5

Histogram of verbal fluency T-score change from T1 to T4 (positive scores refer to a T-score increase over the follow-up period, and negative scores refer to a decline).

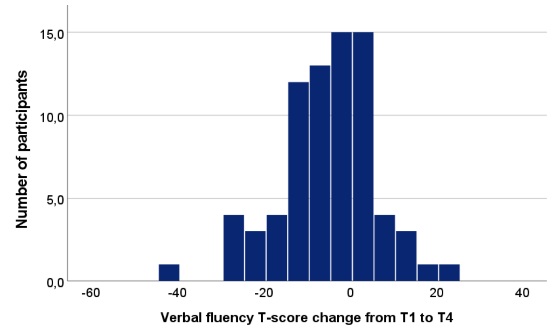


Figure 6

Histogram of visual memory T-score change from T1 to T4 (positive scores refer to a T-score increase over the follow-up period, and negative scores refer to a decline).

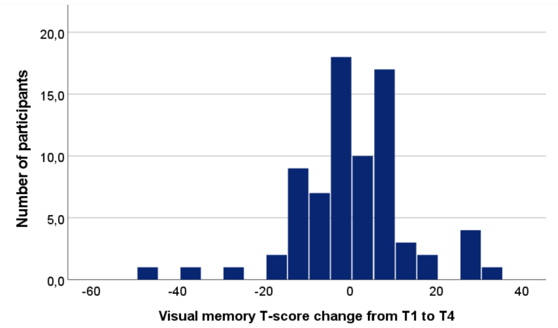


Figure 7

Histogram of verbal memory T-score change from T1 to T4 (positive scores refer to a T-score increase over the follow-up period, and negative scores refer to a decline.

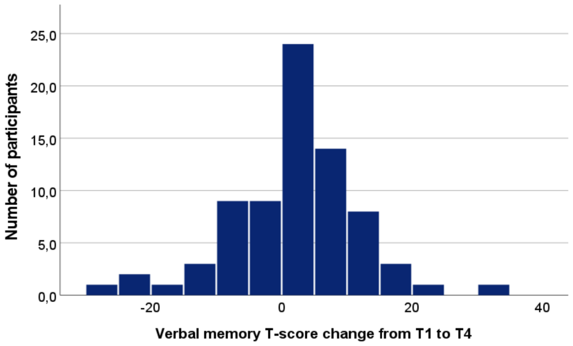
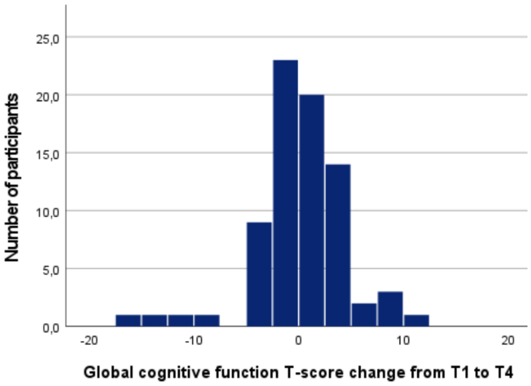


Figure 8

Histogram of GCF T-score change from T1 to T4 (positive scores refer to a T-score increase over the follow-up period, and negative scores refer to a decline).



In Figure 9 we can see that there are only two participants who meet the criterion for PS dysfunction at both T1 and T4. There are five participants who do not meet the criterion for PS dysfunction at T1, but who do so at T4. However, surprisingly, there are six participants who meet the criterion for PS dysfunction at T1, but don’t meet this criterion at T4. 63 participants did not meet the criteria for PS dysfunction at either T1 or T4, these were excluded from the bar graph to make the graph more readable. In Figure 10 we can see that four participants are categorized with both PS dysfunction at T1 and cognitive dysfunction at T4. Seven participants were categorized with intact PS at T1 but cognitive dysfunction at T4, and four participants were categorized with PS dysfunction at T1 and intact cognition at T4. 61 participants had both intact PS at T1 and intact cognition at T4 and were excluded from the graph for readability.

Figure 9

Bar chart of different combinations of PS dysfunction at T1 and T4 (63 participants were categorized with non-dysfunctional PS at both T1+T4).

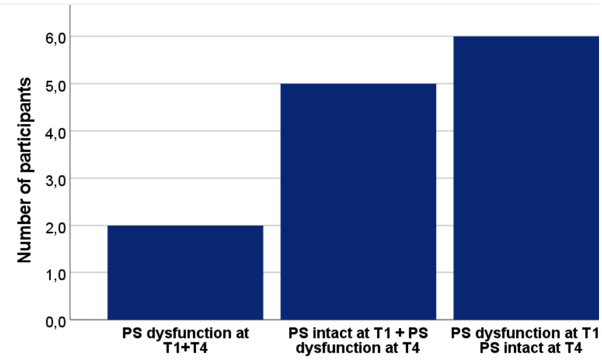
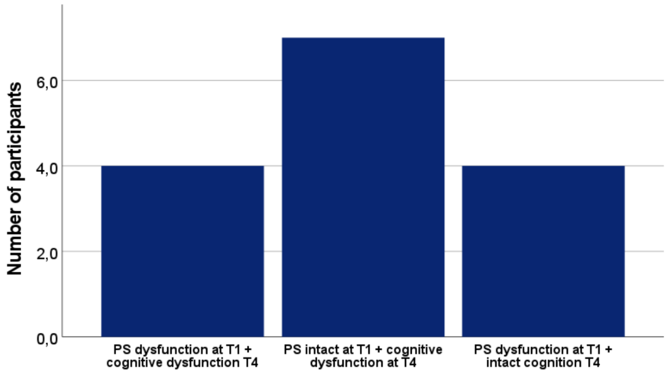


Figure 10

Bar chart of different combinations of T1 PS dysfunction and T4 cognitive dysfunction (61 participants were categorized with both intact PS at T1 and intact cognition at T4).



3.2 Statistical analyses

3.2.1 Missing data

Before performing any analysis of our data, we had to deal with a critical factor in longitudinal studies; participants who dropped out from the study for various reasons. We had data for most of the participants for the first and second test administrations. For the fourth administration, we had data for 55 out of the 76 original participants. Using a complete case analysis would mean that we had to delete 21 rows of participants. Deleting all these rows would make our analysis lose statistical power and assume that the data were missing completely at random, which may increase bias (Sinharay, Stern & Russel, 2001). To maintain statistical power and minimize bias, we decided to use an imputation technique to estimate the missing data points.

We chose to use the Expectation-maximization (EM) algorithm to estimate these data points, a maximum likelihood model. This model has two steps. In the “M” step, the algorithm performs a maximum likelihood analysis as if there were no missing data. The “E” step looks at the expected missing values based on the data and the current parameters, then replaces the missing data with estimated ones. This comparison is made until convergence between the two (E and M) distributions is reached. When this was complete, we filled in the dataset gaps with the values from the generated dataset (Little & Rubin, 1989). This model will assume that the data are missing at random. If a variable can explain the pattern of dropout, this will bias our model and our imputed values.

3.2.2 The relationship between processing speed at T1 and global cognitive function at T4

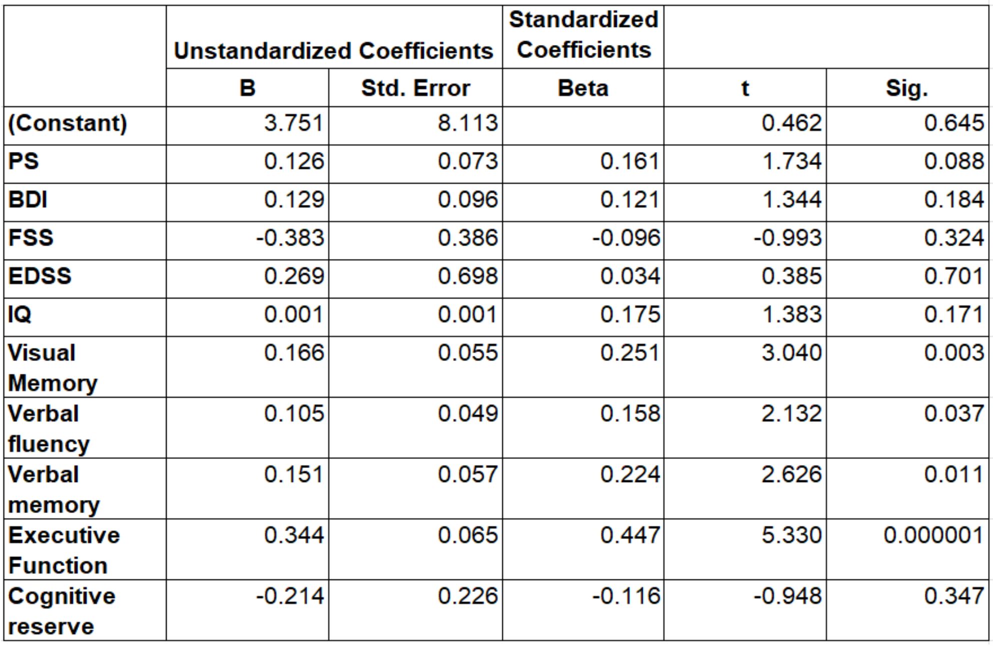
The first linear regression we performed was an analysis using PS at T1 as the predictor and GCF at T4 as the dependent variable, showing a statistically significant relationship with a slope coefficient (the unstandardized beta coefficient) of 0.459. The slope coefficient means that for every unit score of the predictor variable, the predicted score of the outcome variable increases with the value of the slope coefficient. In this example the slope coefficient means that the predicted GCF T-score at T4 increases by 0.459 for each PS T-score point at T1. The relationship between baseline PS and GCF at T4 translates to a Cohen’s *d* of 1.450, which by convention is considered a large effect size (Fritz, Morris & Richler, 2012).

We then performed a multiple regression analysis to control for some known confounding variables. The variables controlled for were BDI scores to control for depression, FSS scores to control for fatigue, EDSS scores to control for disease progression, IQ, verbal memory, verbal fluency, visual memory, and EF at T1 to control for other baseline cognitive data. The relationship between PS at T1 and GCF at T4 was statistically significant; however, the slope coefficient decreased to 0.144. Other statistically significant predictors were visual memory, verbal fluency, verbal memory, and EF with slope coefficients ranging from 0.106 to 0.330.

We performed the same analysis as the latter mentioned, but we also included cognitive reserve as a predictor to control for this (see Table 5). In this analysis, PS at T1 was no longer statistically significant. However, cognitive reserve was not statistically significantly related to GCF at T4 either. When cognitive reserve is included in the model, all cognitive domains except for PS are statistically significant, with EF having the biggest slope coefficient of 0.344 of all the significant predictors.

Table 5

Multiple linear regression analysis using PS at T1 as predictor and GCF at T4 as dependent variable, controlling for some known confounding variables.



	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	3.751	8.113		0.462	0.645
PS	0.126	0.073	0.161	1.734	0.088
BDI	0.129	0.096	0.121	1.344	0.184
FSS	-0.383	0.386	-0.096	-0.993	0.324
EDSS	0.269	0.698	0.034	0.385	0.701
IQ	0.001	0.001	0.175	1.383	0.171
Visual Memory	0.166	0.055	0.251	3.040	0.003
Verbal fluency	0.105	0.049	0.158	2.132	0.037
Verbal memory	0.151	0.057	0.224	2.626	0.011
Executive Function	0.344	0.065	0.447	5.330	0.000001
Cognitive reserve	-0.214	0.226	-0.116	-0.948	0.347

3.2.3 The relationship between PS at T1 and cognitive decline at T4

A binary logistic regression analysis using PS at T1 as predictor and our dummy coded variable for cognitive decline results in a non-significant relationship. A linear regression analysis using PS at T1 as a predictor and our variable for the T-score change in GCF from T1 to T4 also result in a non-significant relationship. We also performed a test of independence chi-square test using the two dummy coded variables PS dysfunction at T1 and our cognitive decline variable from T1 to T4. The Pearson Chi-Square for this analysis was not statistically significant.

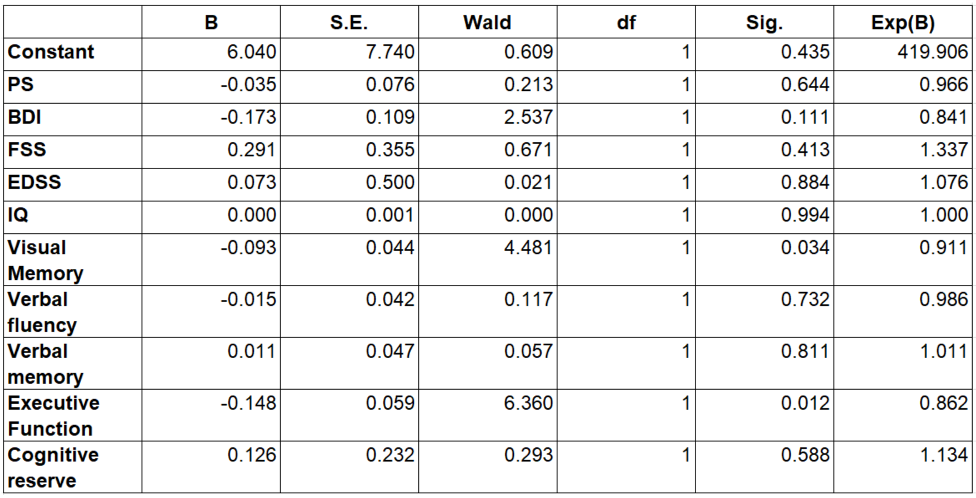
3.2.4 The relationship between PS at T1 and cognitive dysfunction at T4

To explore the relationship between baseline PS and cognitive dysfunction, we performed a binary logistic regression analysis using PS at T1 as a predictor and cognitive dysfunction at T4 as the outcome variable. This analysis resulted in a statistically significant relationship with a slope coefficient of -0.097. In this analysis, a negative slope coefficient means that a higher PS score at T1 is associated with intact cognition at T4. The relationship between baseline PS and cognitive dysfunction at T4 translates to a Cohen's *d* of -0.519. As the effect size is practically on the border of what one conventionally refers to as small and moderate effect size, we choose to call it a "small to moderate" effect size. We then performed the same analysis but controlled for the known confounding variables BDI, FSS, EDSS, IQ, verbal memory, verbal fluency, visual memory, and EF. The results were that PS was no longer statistically significant, but visual memory and EF were.

We then performed the same analysis and controlled for cognitive reserve as a predictor (see Table 6). The results were essentially the same, with PS not being statistically significant and visual memory and EF being statistically significant, with EF having the highest slope coefficient value of -0.148. In this case, a negative slope coefficient also means that higher EF scores are related to intact cognition. Cognitive reserve was not statistically significantly related to cognitive dysfunction at T4.

Table 6

Multiple logistic regression analysis using PS at T1 as predictor and cognitive dysfunction at T4 as dependent variable, controlling for some known confounding variables.



	B	S.E.	Wald	df	Sig.	Exp(B)
Constant	6.040	7.740	0.609	1	0.435	419.906
PS	-0.035	0.076	0.213	1	0.644	0.966
BDI	-0.173	0.109	2.537	1	0.111	0.841
FSS	0.291	0.355	0.671	1	0.413	1.337
EDSS	0.073	0.500	0.021	1	0.884	1.076
IQ	0.000	0.001	0.000	1	0.994	1.000
Visual Memory	-0.093	0.044	4.481	1	0.034	0.911
Verbal fluency	-0.015	0.042	0.117	1	0.732	0.986
Verbal memory	0.011	0.047	0.057	1	0.811	1.011
Executive Function	-0.148	0.059	6.360	1	0.012	0.862
Cognitive reserve	0.126	0.232	0.293	1	0.588	1.134

We performed a test of independence chi square test using the two dummy coded variables PS dysfunction at T1 and cognitive dysfunction at T4 to see whether PS dysfunction at T1 can predict cognitive dysfunction at T4. The Pearson Chi-Square test was statistically significant with a Phi value of 0.346 (see Table 7) and a calculated Cohen's *d* value of 0.738, which is conventionally considered as a moderate effect size.

Table 7

Tests of independence chi square test using PS dysfunction at T1 and cognitive dysfunction at T4 .

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	9.116	1	0.003		
Continuity Correction	6.191	1	0.013		
Likelihood Ratio	6.673	1	0.010		
Fisher's Exact Test				0.013	0.013
Linear-by-Linear Association	8.996	1	0.003		
N of Valid Cases	76				

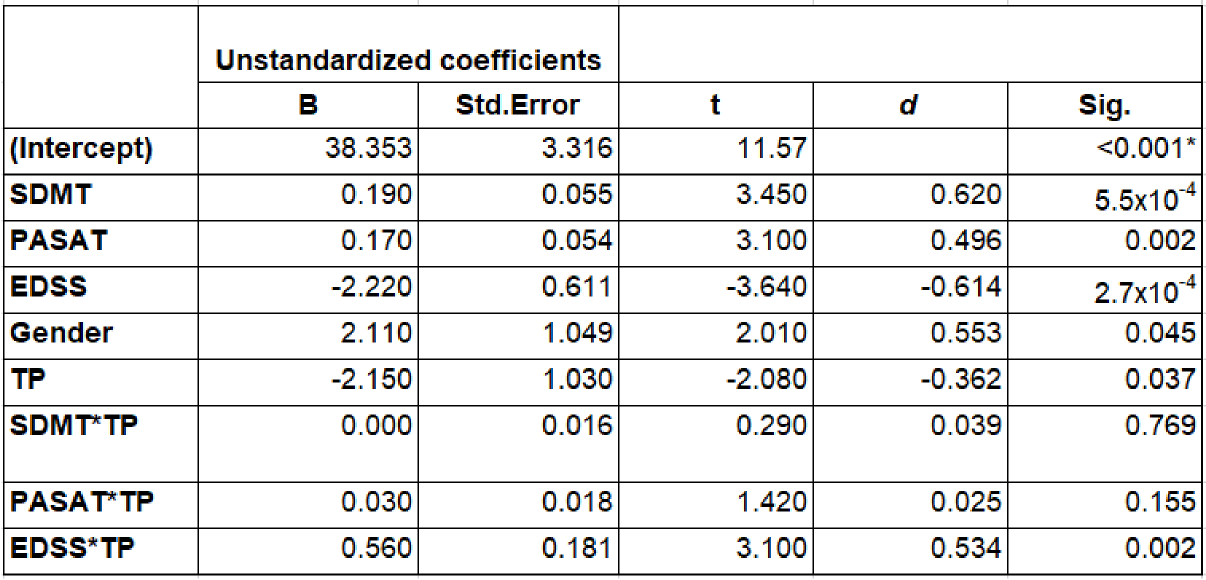
3.2.5 Linear mixed effects models approach to the relationship between baseline PS and GCF at T4

To assess the longitudinal relationship between PS at all TPs and GCF, we used PASAT and SDMT as predictors and GCF as the outcome variable. This analysis was done using an LME model. We controlled for EDSS, gender, and TP. Using the robust version, this model ended up with *p*-values below .05 on all variables except for the interaction of time and SDMT and time and PASAT (see Table 8). The slope coefficient of SDMT in this model is 0.19, and 0.17 for the PASAT. The model ended up with a marginal R^2 of 0.500 and a conditional R^2 of 0.861. The marginal R^2 is interpreted as a normal R^2, meaning the variance is explained by the fixed factors of the model. In this case, fixed factors refer to the predictor variables used in the analysis. The conditional R^2 is the variance explained by both the fixed and random factors in the model.

The data for SDMT as a predictor gives us a Cohen’s *d* of 0.620, which translates to a moderate effect size. The data for PASAT gives us a Cohen’s *d* of 0.496, which translates to a small to moderate effect size. We also performed this analysis using a model where PS replaced SDMT and PASAT. In this model, PS had a slope coefficient of 0.22 and a marginal *R*^2 of 0.116. We did the same analysis using the normal LME function as well. The results were essentially the same, with the only notable difference being a slightly larger marginal R^2 of 0.552 when using SDMT and PASAT as predictors separately in the same model.

Table 8

Robust LME-analysis using SDMT and PASAT at T1 as predictors and GCF at T4 as the dependent variable, including some known confounding variables.



	Unstandardized coefficients				
	B	**Std.Error**	**t**	***d***	**Sig.**
(Intercept)	38.353	3.316	11.57		<0.001*
SDMT	0.190	0.055	3.450	0.620	$5.5x10^{-4}$
PASAT	0.170	0.054	3.100	0.496	0.002
EDSS	-2.220	0.611	-3.640	-0.614	$2.7x10^{-4}$
Gender	2.110	1.049	2.010	0.553	0.045
TP	-2.150	1.030	-2.080	-0.362	0.037
SDMT*TP	0.000	0.016	0.290	0.039	0.769
PASAT*TP	0.030	0.018	1.420	0.025	0.155
EDSS*TP	0.560	0.181	3.100	0.534	0.002

*Our LME model only gives 25 decimals, this means that the p-value of the intercept is less than $1x10^{-25}$

3.2.6 Supplementary analyses

We ran a linear regression analysis to investigate the relationship between SDMT at T1 and GCF at T4. This resulted in a statistically significant relationship with a slope coefficient of 0.356 and a Cohen's *d* of 1.189. We then ran the same analysis using PASAT at T1 as the predictor. These results were also significant, with a slope coefficient of 0.337 and a Cohen's *d* of 1.139.

We ran a binary logistic regression analysis to investigate the relationship between SDMT at T1 and cognitive dysfunction at T4. The result is a statistically significant relationship, with a slope coefficient of 0.041. We then ran the same analysis using PASAT at T1 as the predictor. The results were, however, not statistically significant.

Next, we ran a linear regression analysis using cognitive reserve as a predictor and GCF at T4 as the dependent variable. This analysis resulted in a statistically significant relationship. Finally, a binary logistic regression analysis was statistically significant using cognitive reserve as a predictor and cognitive dysfunction at T4 as the dependent variable.

3.2.7 Post-hoc analyses investigating the relationship between EF and later cognitive outcome

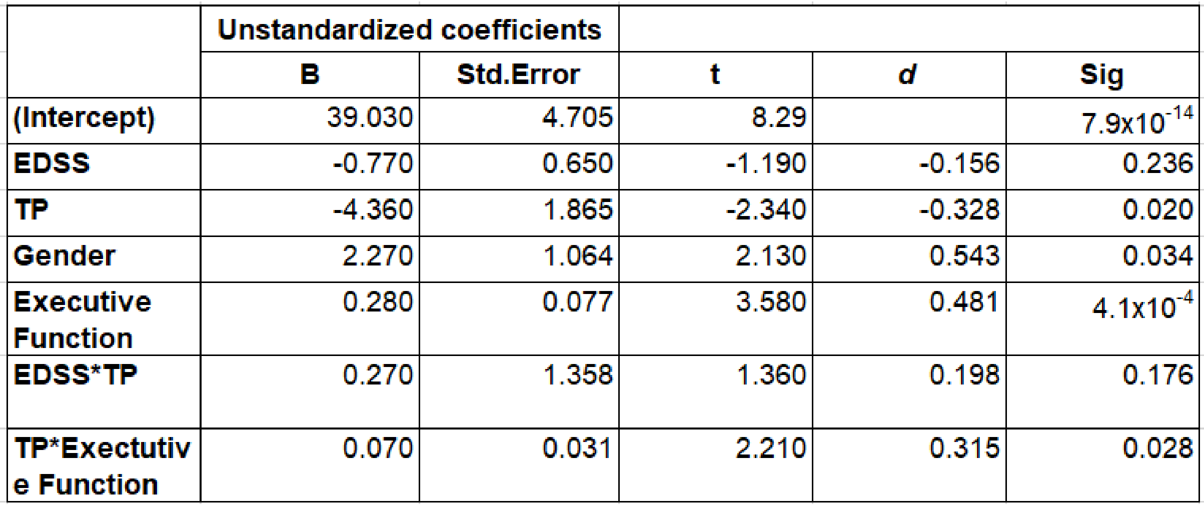
We performed a post-hoc linear regression analysis using only baseline EF as a predictor and GCF at T4 as a dependent variable to investigate whether EF is a better predictor than PS. The results showed that EF was a statistically significant predictor with a slope coefficient of 0.484. The relationship between baseline EF and GCF at T4 translates to a Cohen's *d* of 1.614, which by convention is considered a large effect size and is a slightly larger effect size than the relationship between baseline PS and GCF at T4.

Furthermore, we utilized a post-hoc binary logistic regression analysis, resulting in a non-significant relationship between baseline EF and later cognitive decline. Another post-hoc binary logistic regression analysis using baseline EF as a predictor and cognitive dysfunction at T4 as the outcome variable find a statistically significant relationship with a slope coefficient of -0.121. The relationship between the latter two variables translates to a Cohen's *d* of -1.03.

We also made a model for the repeated measurement longitudinal relationship between baseline EF and GCF at T4 (see Table 9). We controlled for gender, TP, and EDSS in the linear mixed-effects model. This analysis had all variables except EDSS, and EDSS*TP emerge statistically significant. EF had a slope coefficient of 0.28 and a Cohen's *d* of 0.48. The model had a marginal R^2 of 0.305 and a conditional R^2 of 0.702.

Table 9

Robust LME-analysis using baseline EF at T1 as predictor and GCF at T4 as dependent variable, controlling for some known confounding variables.



	Unstandardized coefficients				
	B	Std.Error	t	*d*	Sig
(Intercept)	39.030	4.705	8.29		$7.9x10^{-14}$
EDSS	-0.770	0.650	-1.190	-0.156	0.236
TP	-4.360	1.865	-2.340	-0.328	0.020
Gender	2.270	1.064	2.130	0.543	0.034
Executive Function	0.280	0.077	3.580	0.481	$4.1x10^{-4}$
EDSS*TP	0.270	1.358	1.360	0.198	0.176
TP*Exectutive Function	0.070	0.031	2.210	0.315	0.028

4 Discussion

PS did not predict cognitive decline. However, there was a significant association between baseline PS and GCF at the six-year follow-up visit. This association corresponds to a large effect size, but we could not find this relationship when we controlled for other cognitive domains, IQ, EDSS, FSS, BDI and cognitive reserve.

Baseline PS could predict later cognitive dysfunction, and this relationship corresponds to a small to moderate effect size. However, as with our GCF findings, we could not find this association when we controlled for the above-mentioned confounding variables.

We also found that a dysfunction in baseline PS could predict cognitive dysfunction at the last follow-up visit with a moderate effect size. An interesting finding was a relationship between baseline SDMT and later cognitive dysfunction, but we did not replicate this relationship with the PASAT test. This indicates that SDMT might be a better test for predicting cognitive dysfunction than PASAT. This finding resonates well with our hypothesis that PS is related to later cognitive outcome in people with MS, as performance on SDMT relies heavier on PS than performance on PASAT. This finding also coincides well with other studies investigating cognition in people with MS (see for example Bever, Grattan, Panitch & Johnson, 1995; Fisk & Archibald, 2001; Sonder et al., 2014).

4.1 Why baseline PS can predict cognitive dysfunction but not cognitive decline

We interpret the discrepancy that baseline PS can predict cognitive dysfunction but not cognitive decline as that participants who perform poorly on neuropsychological tests at baseline also perform poorly on neuropsychological tests at follow-ups. This essentially means that these results reflect stability in poor test performance. However, Figure 10 shows the distribution of different combinations of PS dysfunction at T1 and cognitive dysfunction at T4 for our participants. In this bar graph, we can see that only four participants had dysfunction in both PS at T1 and general cognition at T4. There is a broader spread of other combinations of PS dysfunction/intact at T1 and cognitive dysfunction/intact at T4. This means that we cannot find sufficient support for our interpretation of the above-mentioned discrepancy in this figure. Despite this counterargument, we have no better understanding of this discrepancy. Another potential explanation is that we have too few participants who have a cognitive decline over our follow-ups. Of 76 participants, only 11 fall into the category of having a cognitive decline. There may not be enough participants in this subgroup for analyses to reveal any statistically significant relationships.

4.2 SDMT, PASAT, and the operationalization of processing speed

The operationalization of PS was the average score of the SDMT and PASAT. It consisted of two tests because having more data on PS should give us less random error as per the principle of aggregation, increasing the validity of our PS variable. Additionally, SDMT and PASAT scores are moderately to largely correlated at all follow-ups, indicating that they measure the same construct. As described in the Methods section, we know that the psychometrics of both tests are solid. They are both used today in MS-related test batteries, making our results very comparable to other studies.

The administration of PASAT in this study is different from a normal PASAT in that it has a three second delay between numbers as opposed to the standard 1.2s, 1.6s, 2.0s, and 2.4s trials. The use of the three second version compared to the two second version was also recommended by the 1994 National Multiple Sclerosis Society Clinical Outcomes Assessment Task Force (Rudick et al., 1997). It is also comparable to other MS assessment batteries by being administered in the MACFIMS (Benedict et al., 2006) and Brief Repeatable Battery of Neuropsychological Tests (BRB-N; Boringa et al., 2001) batteries. Whereas in the BICAMS battery, SDMT is chosen over PASAT due to its easier administration and less use of equipment (Langdon et al., 2012). In the MACFIMS the PASAT 3" was also found to be the closest related PS measure to vocational disability (Benedict et al., 2006). A good implication of this choice is that it minimizes the risk of using the alternate answers strategy (Tombaugh, 2006). This could be a risk in an MS population due to the expected reduction in PS. On the other hand, it brings negative consequences such as less availability of norms compared to the standard version.

The inclusion of both tests in the operationalization of PS requires focus on some methodical concerns. In a study of long-term results in the PASAT and SDMT as compared to the BRB-N, a significant finding was that the PASAT 3" was affected by a ceiling effect (Sonder et al., 2014), which means that the fact that we used the average of the SMDT and PASAT scores may have biased our results towards minor dysfunction in PS, thus possibly reducing our models predictive ability. A different methodical concern is that the PASAT has a working memory component (Fisk & Archibald, 2001) as you need to keep the last number in your working memory to add the next to it. Performance on PASAT is therefore reliant on working memory in addition to PS, weakening our argument of using an average.

We found that averaging the test scores for our PS variable affects LME model results. Further, the LME models using both SDMT and PASAT as predictors separately performed better than using our averaged PS variable alone as a predictor. However, in the regression analysis, the R^2 remained roughly the same when comparing a model using PS with one using SDMT+PASAT separately. We hypothesize that using the average score of SDMT and PASAT affects the LME model negatively because aggregation reduces the amount of information put into the model (four PS scores versus four SDMT and four PASAT scores). In addition, averaging these scores does not account for error; meanwhile, the LME model accounts for this. One source of error is the methodological concern of the aforementioned ceiling effect limiting the PASAT and these limitations carrying over to the averages we bring into the model.

In the regression analyses, there was a larger effect size for our baseline PS variable compared to only SDMT or only PASAT in relation to GCF at T4, showing that using both tests for prediction is better than either one of them alone. Due to this and the fact that the regression models find roughly the same explained variance using the PS variable and SDMT+PASAT separately as predictors in the same model predicting later cognitive outcome, we argue that both SDMT and PASAT should be included as predictors. Using PS should not necessarily be preferred over SDMT+PASAT separately based upon predictive properties. LME-analyses nuances this relationship by showing that with multiple measurements over time, there is a higher explained variance with SDMT and PASAT as separate predictors, compared to the aggregated PS variable.

To conclude, if prediction of later cognitive outcome based on our models were to be implemented clinically, linear regression analyses are more convenient and clinically suited for prediction compared to LME-analyses, and both SDMT and PASAT should be included as predictors, either aggregated or separately. The latter mentioned analysis requires more advanced statistical knowledge as well as multiple measurements, where the latter is information that will not be available for many patients.

4.3 Is it clinically useful to predict later cognitive function using baseline PS?

Although baseline PS does not predict any later *change* in cognitive function, baseline PS can say something about what cognitive function level we would expect later. For example, if a patient gets a baseline PS T-score of 40, we will also expect a later GCF score in the lower mid-range. Baseline PS is clinically useful because it gives information about current cognitive function. Because cognition generally is stable over time, which is also generally the case for our participant group, it has predictive value because we expect the later cognitive function to be similar to the baseline function. This study does not find indications for clinically meaningful use of baseline PS scores outside of measuring PS as a part of a routine neuropsychological examination.

4.4 Cognitive decline in people with MS

As shown in Figure 2 and Table 4, there is no general tendency of cognitive decline over time in our participant group. On the contrary, there is a general increase in cognitive performance over the first two follow-ups. This might be another explanation for why there is no relationship between baseline PS and cognitive decline. If the cognitive function of our participants doesn't decline over time, there is no decline to predict. This lack of a general cognitive decline tendency in our participant group could not be explained by resilience due to cognitive reserve either. However, there are 11 participants who perform worse on the neuropsychological tests over time, but these cannot be predicted using the methods in this study.

There are few comparable longitudinal studies monitoring cognitive function in people with MS over a long period of time (e.g. 10 years) with multiple follow-ups that control for some relevant confounding variables (Sumowski et al., 2018; Amato, Zipoli & Portaccio, 2006). Some studies that have investigated this find some cognitive decline over follow-ups, however due to the above-mentioned limitations we simply cannot be certain that cognitive function in people with MS generally decline over time. Our findings indicate cognitive stability. It is also possible that six years of follow-up is a too short timespan for detecting cognitive decline in this clinical group, and that a follow-up of for example 10 or 15 years from disease onset could detect cognitive decline.

One possibility is that our participant group is representative of a newer modern MS population in which recently approved medications have a stronger preventative effect on cognitive decline compared to older treatments. At T4, 24 participants were currently using these recently approved medications, while 15 participants had tried these newer medications earlier. If these new medications have a stronger preventative effect compared to older treatment, we expect this effect to be reflected in our participant group, potentially explaining, at least partly, the lack of expected cognitive decline observed in our participants.

If cognitive function in people with MS generally doesn't decline over time, then the question of when cognition first gets impaired in people with MS arises. We know that cognitive impairment is overrepresented in the MS-population compared to the general population (Chiaravalloti & Deluca, 2008; Patti, 2009; Sumowski et al., 2018). One possible explanation is that cognitive impairment occurs more closely in time to the onset of the disease than this study has investigated. Another possibility is that there is a slow decline in cognitive function leading up to the onset of the disease. A third possibility is that cognitive impairment is present from birth. We think that the latter explanation is the least likely of these three. This issue will be up to future studies to investigate.

4.5 Exploring characteristics of participants who had a cognitive decline over the study course

We explored our dataset to investigate if the 11 participants who experienced cognitive decline over our study course had different group characteristics compared to the participants who did not experience cognitive decline. We looked into cognitive reserve, BDI, FSS, EDSS, IQ, SDMT, PASAT, EF, GCF, PS dysfunction, cognitive dysfunction, active medications, medication history, years of education and gender at both T1 and T4. We performed no additional analyses and mainly looked into descriptive information from our data.

One prominent finding is that the mean BDI score of the cognitive decline group at T4 is 15 compared to seven in the cognitively stable group, which is a substantial difference and a staggering mean BDI score increase of 10 compared to T1 for the decline group. Furthermore, the cognitive decline group had a higher mean FSS score, and a lower mean

T-score on SDMT, PASAT, and GCF compared to the cognitively stable group at T4. None of these group differences were prevalent at T1. The only difference found between these two groups at T1 is that the decline group had a lower mean EF T-score compared to the cognitively stable group. As there were no differences between these two groups at T1 except for EF which we had already investigated in our analyses, we found no further information that can be used to predict which participants will have a later cognitive decline by looking more specifically at baseline data of the cognitive decline group. We also found no notable group differences in factors like gender, education, active treatment and treatment history.

The third variable problem initially complicates the interpretation of the BDI finding. It is possible that cognitive decline over time increases depressive symptoms reflected by BDI, however, it is also possible that increased depressive symptoms impairs performance on neuropsychological tests. Other information from our dataset supports the latter direction. If one assumes that cognitive decline in people with MS is a result of disease progression, which is the natural assumption here as we see no other reasonable cause for a real group-level cognitive decline, then a counterargument for the former interpretation is that there are no notable difference in disease progression for the cognitive decline and cognitively stable groups at T4 as their median EDSS scores are equal with similar score distributions. This means that cognitive decline cannot be explained by an increase in disease progression in our dataset. Increased scores for fatigue measured by FSS as well as worse performance on neuropsychological tests related to PS (SDMT and PASAT) are results that coincide well with depressive symptoms (Diamond et al., 2008). Based upon these arguments, we conclude that the cause of cognitive decline in some of our participants is likely due to depressive symptoms and not the MS disease directly. We argue that it is likely that MS is an indirect cause for cognitive decline, as it is a burdensome and serious disease that may contribute to development of depressive symptoms.

4.6 Too strict operationalization of cognitive decline

Our dummy coded cognitive decline variable is, as described earlier, operationalized as a GCF T-score difference of a minimum of -3 from T1 to T4. One could argue that this operationalization is too strict for analyses to reveal statistically significant results. However, there was no relationship between baseline PS and the presumably more sensitive GCF

T-score change from T1 to T4 continuous variable either. As there is no relationship between these latter mentioned variables, the lack of a cognitive decline finding is not due to a too strict operationalization of cognitive decline.

The reasoning behind operationalizing this variable with a T-score decline of three from T1 to T4 instead of a decline of one, was to make our results clinically relevant. One can be more confident that there has been a real decline instead of random measurement errors in GCF over our follow-ups when decline is more strictly defined. Also, a bigger score difference is more clinically meaningful in the way that a bigger decline has a more noticeable impact on the patient's life than for example a GCF score decline of one.

4.7 Executive function and cognitive outcome

PS was not superior to other cognitive domains in the prediction of later cognitive outcome, contrary to what we expected. EF emerges as the baseline predictor with the strongest associations to later cognitive outcome. Baseline EF has a larger effect size in relation to later GCF and cognitive dysfunction than PS. We also found that after controlling for known confounding variables in our analyses, EF was still statistically significantly related to the cognitive outcome variables, which PS was not. This means that if one were to measure a single cognitive domain to predict later cognitive outcome in people with MS, this study indicates that one should measure EF over PS.

Since EF was a predictor that stood out in the regression-based approach. We looked at it as a stand-alone predictor in the LME approach. There was a moderate effect size in relation to GCF at T4, but more interesting was the fact that we saw a drop in the slope coefficient of EDSS compared to the SDMT+PASAT model. The cause of this might be because EF in the LME analysis correlated twice as much with EDSS as PASAT+SDMT.

We point out that EF also fails to predict later cognitive decline or which participants go from cognitively intact at baseline to meeting the criterion for cognitive dysfunction at the last follow-up. We also point out that the same limitations for interpretation of the association between baseline PS and later cognitive dysfunction applies to EF. That is, we argue that this finding reflects stability in performance on neuropsychological tests over time rather than a change in cognitive function.

We want to add that our analyses investigating baseline EF and later cognitive outcome are post-hoc analyses as they are a result of further exploration of our findings outside of our original research questions. This approach carries some limitations in the interpretation of this finding. As we performed several analyses in this study, an unexpected finding is more prone to being a false positive result due to chance. Since we expected PS to have the strongest associations due to the nature of the MS disease mechanism (demyelination) and how the mechanism resonates with the nature of PS, and the fact that we find another cognitive domain with stronger associations without a similarly logical theoretical basis is an argument for a possible false positive result. A counterargument to a false positive result is that our analyses reveal very small *p*-values in the association between baseline EF scores and later cognitive outcome. Another counterargument is that other studies also find EF as one of the cognitive domains that most often is affected in people with MS (Patti, 2009; Chiaravalloti & Deluca, 2008; Sumowski et al., 2018). However, we still view this result as an exploratory one with accompanying limitations in interpretation and conclude that further research is needed to confirm that EF is a better predictor than PS in relation to later cognitive outcome.

4.8 Explanations for cognitive test score patterns over follow-ups

We see a general increase in neuropsychological test performance over our follow-ups, with the exception of a decline from our second to third follow-up. This is generally the opposite finding of what we expected. One possible explanation for the improved test performance over time is practice effects. If participants remember the gist or parts of a test, it could give them an advantage at a later administration, resulting in improved test scores.

Studies find minimal practice effects for SDMT and small practice effects for PASAT with a longitudinal repeated measurement design for people with MS (Sonder et al., 2014; Bever et al., 1995). A study investigating practice effects for CVLT-II and BVMT-R in people with MS find moderate to large practice effects over a one-week follow-up, but practice effects gets substantially reduced when switching between standard and alternate forms (Benedict, 2005), or with a time period between baseline and retest of over one year (Paolo, Troster & Ryan, 1997). These findings are also supported by Woods et al. (2006) and Benedict & Zgaljardic (1998) for healthy adult populations, however the latter study investigated practice effects in the Hopkins Verbal Learning Test-Revised, which is a verbal memory list learning test similar to CVLT-II. We find no longitudinal repeated measurement studies investigating practice effects in CVLT-II specifically. Studies generally find small practice effects for the FAS condition in COWAT (Beglinger et al., 2005; Zgaljardic & Benedict, 2001; Basso, Bornstein & Lang, 1999), however most of these studies investigate practice effects when switching between alternate forms. Studies investigating practice effects in Stroop find large practice effects over repeated measurements and a small practice effect for one follow-up (Beglinger et al., 2005; Basso et al., 1999).

The practice effects found in the above-mentioned studies support the possibility of practice effects in our participant group. A counterargument to this is that the test performance pattern over time in this study deviates from the expected practice effect pattern derived from research. For example, both SDMT and PASAT have a higher improvement pattern over follow-ups in this study than expected based upon practice effect results of the above-mentioned studies. COWAT scores generally decline over time instead of increasing, and as mentioned, there is a general drop in cognitive test performance from T3 to T4. Other counterarguments to practice effects in our dataset is that in this study, alternate forms of CVLT-II and COWAT were used at T2 as well as alternate forms of BVMT-R at every follow-up to minimize practice effects. Also, the time gaps between neuropsychological testing (one, four and six years from baseline respectively) in our study generally is larger than in studies investigating practice effects in neuropsychological tests, further minimizing practice effects.

There are other factors that may contribute to the performance pattern on the neuropsychological tests in this study. Participants may have learned how to generally perform better in a test situation, e.g., learning how to generally tackle neuropsychological tests by getting repeatedly exposed to the test situation or realizing how straining these tests are and planning their day/sleep/other examinations to be prepared and well rested for the tests. Additionally, these tests have been administered by different people throughout the course of the study and have not been administered at consistent times during the day/week. Finally, at the fourth testing dropout will have the most considerable effect. Regarding the effects of medication, we argue that modern medication does not contribute to or explain *improvement* in cognition as these medications slow disease progression and does not cause healing of the disease, but that modern medication potentially may contribute to a better *preservation* of cognitive function or a slowed cognitive decline due to preventive effects (Tintore et al., 2019).

Another possible explanation for the decline from T3 to T4 could be that this generally is the time it takes before a cognitive decline starts or becomes observable through neuropsychological tests in the disease course. Most of the participants got the MS diagnosis at roughly the same time period (sometime between 2009 and 2012), meaning that they roughly are at the same point in time after diagnosis. If cognitive decline generally becomes observable around five years after disease debut, our cognitive decline results from T3 to T4 could potentially reflect the start of observable cognitive decline. The decline in COWAT scores over time may reflect a real decline in phonetic fluency function over the follow-ups.

We will not conclude upon the reason for our participants performing gradually better on these tests until T3 and then having a drop in function at T4, we simply discuss possible factors involved in this pattern. We do argue that practice effects probably contribute to the test performance pattern, however we don't think that practice effects can explain this pattern entirely on its own.

4.9 The research design

We built our research design around an established prospective MS sample, to be able to detect future changes in cognition. We chose the time intervals between assessments because it is long enough to lessen the impact of practice effects but still short enough to get a decent continuous view of the changes, and that it corresponded to ongoing studies that got funded during the follow-up period.

Instead of simply seeing that there is a difference after six years, an aim was to see where the difference emerged and how it evolved. The cost of this knowledge is evident, both financially and methodologically. The strain of participation over six years could lead to increased dropout. This approach could also increase the impact of practice effects as mentioned above. It also requires lots of test materials, adequate test personnel, and test locations. Despite this, pinpointing the start of cognitive decline could be highly useful in increasing the quality of life for people with MS.

Being a longitudinal study our design has the inherent weakness of not being able to conclude on the matter of causality. However, it does give a more thorough understanding of cognitive development than using a cross-sectional design, and might reveal associations that can be further investigated experimentally to conclude upon causality. The design can control for some confounding variables, but has no solution to the third variable problem. It also faces the directionality problem, which cannot be solved, but we have data indicating the direction of the relationship between our finding of BDI scores and cognitive decline.

4.10 Comparing our results with other studies

The longitudinal perspective of cognitive function in MS consists of a modest body of research. The previous research in this field has had small samples. This makes it more difficult to generalize due to their differences in clinical characteristics, neuropsychological test batteries, definitions of cognitive deficits, and their statistical approach. Some major methodological problems plague these studies, including: samples as mentioned, how to deal with practice effects, differences in duration, and dropout.

One significant difference observed in our design compared to other studies is the lack of a substantial MS participant "impaired" group. In other studies (Strober et al., 2014; Katsari et al., 2020; Sperling et al., 2001), the impaired group makes up everything from 20% to 100% of the group investigated. Strober et al. (2014) mention an increase from 41% to 59% in cognitively impaired people with MS, which is lower than Amato et al., who report an increase from 26% to 56% over 10 years. Schwid et al. (2007) even went as far as to say that the participants that ended up changing groups were the most significant source of increased cognitive impairment in these studies. The number of people making such a qualitative jump in performance was minimal, meaning that the lack of such a group change could partly explain the lack of findings in cognitive decline.

Another significant difference between our study and similar ones is that there was more than one follow-up. For example, Strober et al. (2014) did a baseline assessment and a follow-up at 18 years, as did Katsari et al. (2020), except with ten years. Meanwhile, Amato et al.'s (2006) study design was more similar to ours with a four- and ten-year follow-up. The significant period between each follow-up in some of these studies will naturally make a difference when observing cognitive decline. Another advantage is that the impact of practice effects will naturally be minor with this considerable gap. However, if you take on the risk of practice effects and follow your participants over multiple assessments with shorter time gaps, you are more likely to see the beginning of a trend.

The earlier longitudinal studies looking into PS in people with MS have differing results. Katsari et al. (2020) include a boxplot showing that the performance on PS as measured by SDMT in their follow-up is 0.5 point better for unimpaired participants and 1.5 points worse for impaired participants. In this study, the PASAT 2” and 3” were classified as tests of working memory and not tests of PS, though the score differences on these tests weren’t statistically significant either. This finding is consistent with earlier findings. However, other studies report a statistically significant change in PS over many years (Schwid et al. 2007; Strober et al. 2014). These results make it clear that currently, there is no universal answer to whether PS is constant or deteriorating over the course of the MS disease. The effect of dropout in the aforementioned studies could also cloud the impact of the degree to which the sampling reflects all people with MS. Our study echoes the findings of Katsari et al. in that we do not see a deteriorating performance in measures of PS, even with a decently low dropout rate for six years of following the same group.

As mentioned we see a trend of improvement over the first two follow-ups, and a fascinating small drop in performance at the third follow-up. This drop in performance is consistent with Achiron et al.'s (2013) model.

4.11 Operationalization of GCF, EF and cognitive dysfunction

As described in the method section, our operationalization of GCF is the average T-score of the five cognitive domains PS, verbal memory, visual memory, verbal fluency and EF. We argue it is reasonable to weight them equally to make each cognitive domain contribute equally to this global cognition variable. An alternative way to make this variable could be to make a categorical variable based upon the number of cognitively dysfunctional domains. We considered this, however this operationalization of GCF would make analyses less sensitive to the first research question, which is if processing speed can predict a decline in GCF over time in people with MS. The former operationalization of GCF results in an interval variable and would arguably pick up a more subtle decline in cognition function over time more easily compared to the categorical latter mentioned operationalization.

Utilizing PS as a predictor in most of the analyses and operationalizating GCF in a way that includes PS will cause an increase in the correlation between the predictor and outcome variable as the analyses partly compare PS at T1 with PS at T4. We argue that PS should be a part of the operationalization of GCF due to two factors. The first is that PS is the first cognitive domain expected to be impaired over the time period. The consequences of not including PS in GCF could be that a decline over time in PS may not get picked up in our analyses. The second is that PS is an important component of general cognition. If we exclude PS in the operationalization of GCF, a big chunk of the conceptualization of cognition is left out. We therefore, argue that the benefits of including PS in our operationalization of GCF outweigh the cons.

Regarding the operationalization of EF, the only data we had related to this domain was condition three and four of the D-KEFS Stroop test. Computation of the EF variable was done using the average of the time spent on these two conditions as this was an easy and reasonable way to operationalize EF with the data we had available. Errors made during the Stroop test were not included in this operationalization. This is mainly due to participants generally making few mistakes (e.g., an arithmetic average of roughly 0.5 uncorrected errors and 0.9 self-corrected errors per participant at condition 4 at T4) and the issue of weighting completion time and errors in the operationalization of EF. We argue that completion time is the most important data regarding EF and that errors made are of less importance in this participant group due to the generally low error frequency. We also argue that it would be difficult to include errors in the operationalization of EF in a way that increases construct validity and that a simple yet reasonable solution for this problem is to only use the completion time data for condition three and four.

Clinically cognitive dysfunction is often operationalized as a dysfunction in a minimum of two cognitive domains, in which dysfunction usually is defined by a standardized score of 1.5 standard deviations below the mean which translates to a T-score ≤35 (Lezak et al., 2012, p. 172; Petersen, 2004). To make our results as relevant for the clinician as possible, we decided to use this definition. We also considered operationalizing cognitive dysfunction by a cut-off score of GCF, however, this is simply not how it is done clinically, meaning this operationalization would impact the external validity negatively.

4.12 Participant group characteristics and representativeness

A key premise for our results to be generalized to the general MS population is that our MS participant group reflects the general MS population. Descriptive statistics of neuropsychological test results initially indicate that this might not be the case as the participants as a group scores well above the norm average on multiple tests and at multiple TPs (see Tables 2 and 3 or Table A1), this is probably partly due to these specific tests norms not controlling for years of education. 79% of the participants have completed a bachelor's degree, while only 35.3% of the Norwegian population in 2020 had completed some level of higher education (SSB, 2021). Also, the average IQ of the participant group as measured by WASI is 118.

One reason for a potential deviation from a general Norwegian MS population could be the county where this study took place due to differences in education levels in different geographical locations in Norway. According to "Statistisk sentralbyrå" (SSB), a state statistical organization in Norway, in 2020 53.1% of the inhabitants in the county of "Oslo" had completed education at university level, followed by the county "Viken" with 35% (SSB, 2021). This shows a large discrepancy in education levels between Oslo and other counties in Norway. There are also some district variations in Oslo. This study was carried out at Oslo University Hospital, Ullevål which lies in the district "Nordre Aker". In this district 63.2% of inhabitants have completed education at university level (SSB, 2021). The districts closest to "Nordre Aker" are "Vestre Aker", "St. Hanshaugen", "Sagene" and "Bjerke", with percentages of completed education at university level of 65.3%, 65.0%, 64.8% and 43.7% respectively (SSB, 2021). Vestre Aker was the district with the highest percentage of all the districts in Oslo.

For our MS participant group we have collected data on the highest level of education the participants have completed at T3, and total number of education years at T1. The median education level at T3 was a completed bachelor's degree and 79% of participants have a minimum of a bachelor's degree, which is considerably higher than even the Oslo district with the highest percentage of completed higher education of 65.3%. This was at T3, and the results could be biased by drop-out. At T3 data for updated education level was lacking for 14 participants, however both the arithmetic mean and median level of education of these 14 participants at T1 was 15 years, which is also the average for the entire MS participant group at T1. From this data we can conclude that participants who dropped out did not affect the group proportions of education levels at T3 in a significant way. We can also conclude that the MS participant group has a higher proportion of higher education than the average norwegian, the average Oslo-inhabitant and the average inhabitant of the districts with the highest proportions of higher education in Oslo.

Studies that measure IQ in people with MS find different group average IQ scores. Gotkovsky (2014), Marsh (1980) and Gould et al. (2018) found average IQ scores of 89, 105 and 118 respectively in people with MS. Gould et al. (2018) administered WASI to measure IQ, similar to our study, while Marsh (1980) and Gotkovsky (2014) administered adult versions of the Wechsler intelligence tests. A report by Kunnskapssenteret (Siqveland, Dalsbø, & Harboe, 2014), a state center for health services that convey scientific knowledge about effects of different methods in the Norwegian health service, investigated psychometric properties of WASI for the Norwegian population. They concluded that WASI may overestimate the IQ of Norwegian people because the norms were based on an American population from the 1990's (Siqveland et al., 2014). This could partly explain the measured difference in IQ between the above-mentioned studies as those that use WASI get high average IQ scores. Also, WASI does not measure PS or working memory which are commonly affected cognitive domains in this clinical group (Chiaravalloti & Deluca, 2008), which might additionally increase the discrepancy in IQ scores measured by WASI and WAIS-IV for people with MS.

The treatment of the participant group is a relevant topic of discussion. Most of our participants have gone through different treatments (see Appendix B Figure A22 and A23 for pie charts). The most used either previous or current treatment were interferon alpha and beta. The widespread use of this treatment and its longevity makes our study easier to compare to other studies. Meanwhile, we also have a variety of more modern treatments that makes our study less comparable to previous ones. The use of fingolimod was approved by the FDA in 2010 for the treatment of RRMS (Tintore et al., 2019). The use of teriflunomide was approved by the FDA in 2012 and the use of intravenous alemtuzumab was approved in 2014 (Tintore et al., 2019). The number of participants using these "newer" medications amounts to 24 currently using at T4 and 15 previously having tried them (see Appendix B). While making our study less comparable to previous ones, this will make our study more comparable to newer studies. It will also make our participant group more representative of the current, and newer generations of people with MS with access to these medications. Though how long this argument lasts will depend on the speed of development on the pharmacological side.

The selection process may have contributed to a selection bias in this participant group. As mentioned in the "participants" segment that describes the selection process under "Method", 43 out of 151 patients considered for inclusion to this study were excluded due to fulfillment of at least one exclusion criteria. It is possible that this excluded group has other characteristics than the included group which may have resulted in the included group being biased relative to the general MS population. Furthermore, 85 out of 108 participants were selected due to limits in MRI capacity and nine out of these 85 participants declined. These are all factors that may potentially have contributed to a selected group of participants.

We conclude that our participant group is largely representative for the general modern MS population. Participants were admitted to this study directly from the clinic and not from advertisement, arguably reducing selection bias in this process and increasing external validity. High measured IQ in this study can be at least partly explained by the norm group for the WASI test in relation to the Norwegian population. High consistent scores on some neuropsychological tests can be partly explained by the lack of control for years of education for these specific tests. Lastly, the high average education level might be explained partly by exclusion criteria and which study candidates accept study participation which might have led to a more well-educated participant group, however it is not uncommon for a study to have participants with above-average education.

4.13 Limitations

We operationalized cognitive dysfunction as a T-score of ≤35 on two or more domains to make it generalizable. This made it difficult to place any of our participants in the deficit group. When looking into our first research question it may have been interesting to use a more sensitive operationalization, like a T-score cut-off of 40.

The dropout rate was calculated at 28% for this study which is a limitation, however a good dropout rate considering the duration of the study and compared to other similar studies. We could not find any pattern of characteristics separating those who dropped out from those who stayed. Therefore, the main effects dropout will have on our study is that it will reduce the generalizability, and limit the statistical power.

Furthermore, a six-year follow-up might be too short to uncover cognitive decline. A study by Achiron et al. (2013) suggests that cognitive decline in people with MS becomes observable through cognitive tests after five years of disease duration. Our findings coincide with this as we find a cognitive decline from T3 to T4, which are assessments at roughly four and six years from disease debut respectively. A limitation in our study is that based upon Achiron et al. one should expect further cognitive decline after our T4 measurement, which is a longitudinal aspect of cognition in people with MS that we were unable to investigate.

Every participant was subject to current or previous treatment during the study. We have made an effort to identify which treatment the participant received at different points in time. The limitation here is that it would be challenging to calculate or assess what effect the different treatments could have on the participants due to the vast variation in participants' medication history. The effect of different treatments would be an interesting variable to investigate, but was beyond the scope of our dataset.

All the test norms control for the participants' age. However, different neuropsychological tests have norms based on additional, and different demographic, and educational variables. The gender proportion is skewed towards females and the average years of education is considerably higher than the Norwegian average, and presumably considerably higher than most norm populations. This might bias some standardized neuropsychological test results. For example, BVMT-R and D-KEFS Stroop norms do not control for gender. If there are gender differences in performance on these tests, the average standardized result will be skewed towards the average female performance.

SDMT, PASAT, COWAT and D-KEFS Stroop norms control for education, while BVMT-R and CVLT-II norms don't. This might result in higher standardized scores for BVMT-R and CVLT-II for this highly educated participant group relative to SDMT, PASAT, COWAT, and Stroop because the aforementioned tests don't control for education. The average T-scores for all participants across all TPs for PS, verbal fluency, EF, visual memory and verbal memory are roughly 47, 49, 53, 56 and 61 respectively. The highest average scores are visual memory and verbal memory, and the tests used to measure these domains are BVMT-R and CVLT-II. Meaning education level probably have an effect on the standardized performance on the two latter mentioned tests.

As mentioned, we used Portuguese norms from 2018 (Sousa et al., 2018) as there are few available norms for the three second version of PASAT. Brazilian norms (Damasceno et al., 2018) were an option for the three second version, however we chose the Portuguese norms assuming that the Portuguese population more closely resembles the Norwegian population. Using portuguese norms may have resulted in different standardized scores compared to using Norwegian norms. However, we argue that using the Portuguese norms is better than the alternative, which is not standardizing this PASAT test.

EF is an umbrella term typically described as higher-level cognition involving mainly top-down cognitive processes such as mental flexibility, inhibition, planning, problem solving, attention, working memory and more (Gilbert & Gurbess, 2008). Operationalization of EF as the standardized average time spent on condition three and four on the D-KEFS Stroop test is mainly a measure of inhibition and mental flexibility. Due to the narrow available data we have for EF compared to the wide EF term, a limitation is that our EF variable only reflects a segment of the wider conceptualization of EF.

A weakness in our verbal memory data is the lack of administration of the CVLT-II long-delay recall condition. The consequence of this is that our "verbal memory" variable is solely based upon the short-delay recall condition, limiting the construct validity of this variable.

4.14 Implications for future research

This study has used performance on tests of cognitive function to predict later cognitive decline and dysfunction. In the process of doing so there have been factors found to be of use to highlight for future research. The first factor being that our findings indicate that newly diagnosed people with MS as a group generally appear to be cognitively well functioning. Furthermore, over a six year follow-up we find a high degree of cognitive stability. For future studies it would be interesting to attempt to predict cognitive decline over a longer follow-up period as the onset of cognitive decline might occur later than what this study could investigate.

There is a relationship between cognitive dysfunction and baseline PS. However, as participants largely are cognitively stable over our follow-ups, we were wondering if our design could be improved to deal with some important factors. Using the dataset we procured descriptive information about the different treatments and changes that were made during the study (see appendix B). There was however, not a clear way to split them into groups based on treatment, due to the aforementioned changes over the six year period. Therefore, there was no measure of the effects different treatments had on the participants, only descriptive data. Exploring such effects and investigating their impact on MS cognitive progression could be important to unravel the longitudinal effects MS has on a person using modern medications. Especially considering the emergence of newer and more effective MS treatments (Hauser & Cree, 2020).

Furthermore, we are left wondering if our participants are a highly functioning group of individuals with MS, or if they are a representation of the modern person with MS. Over the last two decades in an area close to where we aquired our participants, one can see a trend of earlier diagnosis and a less severe disease course (Simonsen et al., 2021). This less severe disease course was linked to a longer time to reach EDSS score of six, which may again be linked to our trend of cognitive stability. However, reaching a definite conclusion on this matter is beyond the scope of this thesis.

5 Conclusion

Our participants mainly exhibited cognitive stability over the course of six years. Baseline processing speed was not related to later cognitive decline. However, baseline PS predicted later cognitive dysfunction. For future studies we recommend implementing measures accounting for and reducing the practice related effects on the cognitive tests, by having a control group and longer intervals between assessments. We would also encourage the using cognitive tests with accessible and well-established norms. Further longitudinal research on early predictors of cognitive function in people with MS will be of great benefit to the MS community both in terms of research and patient information. The knowledge obtained could potentially derive a more in-depth understanding of the disease mechanisms and may be used to predict, treat, and perhaps prevent decline in cognitive function in people with MS. The ultimate goal will be to improve the quality of life for people with MS.

References

Achiron, A., Chapman, J., Magalashvili, D., Dolev, M., Lavie, M., Bercovich, E., Polliack, M., Doniger, G. M., Stern, Y., Khilkevich, O., Menascu, S., Hararai, G., Gurevich, M., & Barak, Y. (2013). Modeling of cognitive impairment by disease duration in multiple sclerosis: a cross-sectional study. *PLoS One, 8*(8), e71058. https://doi.org/10.1371/journal.pone.0071058

Amato, M. P., Zipoli, V., & Portaccio, E. (2006). Multiple sclerosis-related cognitive changes: a review of cross-sectional and longitudinal studies. *J Neurol Sci, 245*(1-2), 41-46. https://doi.org/10.1016/j.jns.2005.08.019

Basso, M. R., Bornstein, R. A., & Lang, J. M. (1999). Practice effects on commonly used measures of executive function across twelve months. *Clin Neuropsychol, 13*(3), 283-292. https://doi.org/10.1076/clin.13.3.283.1743

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., Fastenau, P. S., & Siemers, E. R. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Arch Clin Neuropsychol, 20*(4), 517-529. https://doi.org/10.1016/j.acn.2004.12.003

Benedict, R. H. (2005). Effects of using same- versus alternate-form memory tests during short-interval repeated assessments in multiple sclerosis. *J Int Neuropsychol Soc, 11*(6), 727-736. https://doi.org/10.1017/S1355617705050782

Benedict, R. H. B., Cookfair, D., Gavett, R., Gunther, M., Munschauer, F., Garg, N., & Weinstock-Guttman, B. (2006). Validity of the minimal assessment of cognitive function in multiple sclerosis (MACHMS). *Journal of the International Neuropsychological Society*, *12*(4), 549-558. https://doi.org/10.1017/S1355617706060723

Benedict, R. H., Fischer, J. S., Archibald, C. J., Arnett, P. A., Beatty, W. W., Bobholz, J., Chelune, G. J., Fisk, J. D., Langdon, D. W., Caruso, L., Foley, F., LaRocca, N. G., Vowels, L., Weinstein, A., DeLuca, J., Rao, S. M., & Munschauer, F. (2002). Minimal neuropsychological assessment of MS patients: a consensus approach. *Clin Neuropsychol, 16*(3), 381-397. https://doi.org/10.1076/clin.16.3.381.13859

Benedict, R. H., Morrow, S. A., Weinstock Guttman, B., Cookfair, D., & Schretlen, D. J. (2010). Cognitive reserve moderates decline in information processing speed in multiple sclerosis patients. *J Int Neuropsychol Soc, 16*(5), 829-835. https://doi.org/10.1017/S1355617710000688

Benedict, R. H., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *J Clin Exp Neuropsychol*, *20*(3), 339-352. https://doi.org/10.1076/jcen.20.3.339.822

Benedict, R. H. B., Schretlen, D., Groninger, L., Dobraski, M., & Shpritz, B. (1996). Revision of the Brief Visuospatial Memory Test: Studies of normal performance, reliability, and validity. *Psychological Assessment, 8*(2), 145–153. https://doi.org/10.1037/1040-3590.8.2.145

Berg-Hansen, P., Moen, S. M., Harbo, H. F., & Celius, E. G. (2014). High prevalence and no latitude gradient of multiple sclerosis in Norway. *Mult Scler, 20*(13), 1780-1782. https://doi.org/10.1177/1352458514525871

Bergendal, G., Fredrikson, S., & Almkvist, O. (2007). Selective decline in information processing in subgroups of multiple sclerosis: an 8-year longitudinal study. *Eur Neurol, 57*(4), 193-202. https://doi.org/10.1159/000099158

Bever, C. T., Jr., Grattan, L., Panitch, H. S., & Johnson, K. P. (1995). The Brief Repeatable Battery of Neuropsychological Tests for Multiple Sclerosis: a preliminary serial study. *Mult Scler, 1*(3), 165-169. https://doi.org/10.1177/135245859500100306

Bjornevik, K., Cortese, M., Healy, B. C., Kuhle, J., Mina, M. J., Leng, Y., Elledge, S. J., Niebuhr, D. W., Scher, A. I., Munger, K. L., & Ascherio, A. (2022). Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science, 375*(6578), 296-301. https://doi.org/10.1126/science.abj8222

Boringa, J. B., Lazeron, R. H., Reuling, I. E., Ader, H. J., Pfennings, L., Lindeboom, J., de Sonneville, L. M., Kalkers, N. F., & Polman, C. H. (2001). The brief repeatable battery of neuropsychological tests: normative values allow application in multiple sclerosis clinical practice. *Mult Scler, 7*(4), 263-267. https://doi.org/10.1177/135245850100700409

Brown, V. A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science, 4*(1), 2515245920960351. doi:10.1177/2515245920960351

Chen, M. H., Chiaravalloti, N. D., Genova, H. M., & Costa, S. L. (2020).Visual and motor confounds on the symbol digit modalities test. *Mult Scler Relat Disord, 45*, 102436. https://doi.org/10.1016/j.msard.2020.102436

Chiaravalloti, N. D., & DeLuca, J. (2008). Cognitive impairment in multiple sclerosis. *Lancet Neurol, 7*(12), 1139-1151. https://doi.org/10.1016/S1474-4422(08)70259-X

Compston, A., & Coles, A. (2008). Multiple sclerosis. *Lancet, 372*(9648), 1502-1517. https://doi.org/10.1016/S0140-6736(08)61620-7

Costa, S. L., Genova, H. M., DeLuca, J., & Chiaravalloti, N. D. (2017). Information processing speed in multiple sclerosis: Past, present, and future. *Mult Scler, 23*(6), 772-789. https://doi.org/10.1177/1352458516645869

Damasceno, A., Amaral, J., Barreira, A. A., Becker, J., Callegaro, D., Campanholo, K. R., Damasceno, L. A., Diniz, D. S., Fragoso, Y. D., Franco, P. S., Finkelsztejn, A., Jorge, F. M. H., Lana-Peixoto, M. A., Matta, A., Mendonca, A. C. R., Noal, J., Paes, R. A., Papais-Alvarenga, R. M., Pereira, A. G., . . . Damasceno, B. P. (2018). Normative values of the Brief Repeatable Battery of Neuropsychological Tests in a Brazilian population sample: discrete and regression-based norms. *Arq Neuropsiquiatr, 76*(3), 163-169. https://doi.org/10.1590/0004-282x20180006

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987-2000). California Verbal Learning Test--Second Edition (CVLT –II) [Database record]. APA PsycTests. https://doi.org/10.1037/t15072-000

Demaree, H. A., DeLuca, J., Gaudino, E. A., & Diamond, B. J. (1999). Speed of information processing as a key deficit in multiple sclerosis: implications for rehabilitation. *J Neurol Neurosurg Psychiatry, 67*(5), 661-663. https://doi.org/10.1136/jnnp.67.5.661

Diamond, B. J., Johnson, S. K., Kaufman, M., & Graves, L. (2008). Relationships between information processing, depression, fatigue and cognition in multiple sclerosis. *Arch Clin Neuropsychol, 23*(2), 189-199. https://doi.org/10.1016/j.acn.2007.10.002

Dobson, R., & Giovannoni, G. (2019). Multiple sclerosis - a review. *Eur J Neurol, 26*(1), 27-40. https://doi.org/10.1111/ene.13819

Doshi, A., & Chataway, J. (2017). Multiple sclerosis, a treatable disease. *Clin Med (Lond), 17*(6), 530-536. https://doi.org/10.7861/clinmedicine.17-6-530

Egeland, J., Landro, N. I., Tjemsland, E., & Walbaekken, K. (2006). Norwegian norms and factor-structure of phonemic and semantic word list generation. *Clin Neuropsychol, 20*(4), 716-728. https://doi.org/10.1080/13854040500351008

Fisk, J. D., & Archibald, C. J. (2001). Limitations of the Paced Auditory Serial Addition Test as a measure of working memory in patients with multiple sclerosis. *Journal of the International Neuropsychological Society*, *7*(3), 363-372. https://doi.org/10.1017/s1355617701733103

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *J Exp Psychol Gen, 141*(1), 2-18. https://doi.org/10.1037/a0024338

Gilbert, S. J., & Burgess, P. W. (2008). Executive function. *Curr Biol, 18*(3), R110-114. https://doi.org/10.1016/j.cub.2007.12.014

Gontkovsky, S. T. (2014). Influence of IQ in interpreting MMSE scores in patients with multiple sclerosis. *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn, 21*(2), 214-221. https://doi.org/10.1080/13825585.2013.795515

Gould, J. R., Reineberg, A. E., Cleland, B. T., Knoblauch, K. E., Clinton, G. K., Banich, M. T., Corboy, J. R., & Enoka, R. M. (2018). Adjustments in Torque Steadiness During Fatiguing Contractions Are Inversely Correlated With IQ in Persons With Multiple Sclerosis. *Front Physiol, 9*, 1404. https://doi.org/10.3389/fphys.2018.01404

Harrison, J. E., Buxton, P., Husain, M., & Wise, R. (2000). Short test of semantic and phonological fluency: Normal performance, validity and test-retest reliability. *British Journal of Clinical Psychology*, 39, 181-191. https://doi.org/10.1348/014466500163202

Hauser, S. L., & Cree, B. A. C. (2020). Treatment of Multiple Sclerosis: A Review. Am J Med, 133(12), 1380-1390 e1382. https://doi.org/10.1016/j.amjmed.2020.05.049

Hechenberger, S., Helmlinger, B., Ropele, S., Pirpamer, L., Bachmaier, G., Damulina, A., . . Pinter, D. (2022). Information processing speed as a prognostic marker of physical impairment and progression in patients with multiple sclerosis. *Mult Scler Relat Disord*, *57*, 103353. https://doi.org/10.1016/j.msard.2021.103353

Helsedirektoratet. (2021). *Nasjonal faglig retningslinje for diagnostikk, attakk- og* *sykdomsmodifiserende behandling av multippel sklerose.* Obtained from https://www.helsedirektoratet.no/retningslinjer/multippel-sklerose

Homack, S., Lee, D., & Riccio, C.A. (2005) Test Review: Delis-Kaplan Executive Function System, *Journal of Clinical and Experimental Neuropsychology, 27:*5, 599-609, https://doi.org/10.1080/13803390490918444

Høgestøl, E. A. (2020). *MRI and Other Biomakers in Early MS* [Doctoral dissertation, University of Oslo]. DUO Vitenarkiv. http://urn.nb.no/URN:NBN:no-83923

Jennekens-Schinkel, A., Laboyrie, P. M., Lanser, J. B., & van der Velde, E. A. (1990). Cognition in patients with multiple sclerosis After four years. *J Neurol Sci, 99*(2-3), 229-247. https://doi.org/10.1016/0022-510x(90)90158-j

Katsari, M., Kasselimis, D. S., Giogkaraki, E., Breza, M., Evangelopoulos, M. E., Anagnostouli, M., Andreadou, E., Kilidireas, C., Hotary, A., Zalonis, I., Koutsis, G., & Potagas, C. (2020). A longitudinal study of cognitive function in multiple sclerosis: is decline inevitable? *J Neurol, 267*(5), 1464-1475. https://doi.org/10.1007/s00415-020-09720-8

Kingwell, E., van der Kop, M., Zhao, Y., Shirani, A., Zhu, F., Oger, J., & Tremlett, H. (2012). Relative mortality and survival in multiple sclerosis: findings from British Columbia, Canada. *J Neurol Neurosurg Psychiatry, 83*(1), 61-66. https://doi.org/10.1136/jnnp-2011-300616

Kister, I., Bacon, T. E., Chamot, E., Salter, A. R., Cutter, G. R., Kalina, J. T., & Herbert, J. (2013). Natural history of multiple sclerosis symptoms. *Int J MS Care, 15*(3), 146-158. https://doi.org/10.7224/1537-2073.2012-053

Klineova, S., & Lublin, F. D. (2018). Clinical Course of Multiple Sclerosis. *Cold Spring Harb Perspect Med, 8*(9). https://doi.org/10.1101/cshperspect.a028928

Koller, M. (2016). robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. *Journal of Statistical Software, 75*(6), 1-24. https://doi.org/10.18637/jss.v075.i06

Langdon, D. W., Amato, M. P., Boringa, J., Brochet, B., Foley, F., Fredrikson, S., Hamalainen, P., Hartung, H. P., Krupp, L., Penner, I. K., Reder, A. T., & Benedict, R. H. (2012). Recommendations for a Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS). *Mult Scler, 18*(6), 891-898. https://doi.org/10.1177/1352458511431076

Lezak, M.D., Howieson, D.B., Bigler, E.D., Tranel, D. (2012). *Neuropsychological Assessment, Fifth Edition*. New York: Oxford University Press

Little, R. J. A., & Rubin, D. B. (1989). The Analysis of Social-Science Data with Missing Values. Sociological Methods & Research, 18(2-3), 292-326. https://doi.org/10.1177%2F0049124189018002004

Lunde, H. M. B., Assmus, J., Myhr, K. M., Bo, L., & Grytten, N. (2017). Survival and cause of death in multiple sclerosis: a 60-year longitudinal population study. *J Neurol Neurosurg Psychiatry, 88*(8), 621-625. https://doi.org/10.1136/jnnp-2016-315238

Marsh, G. G. (1980). Disability and intellectual function in multiple sclerosis patients. *J Nerv Ment Dis, 168*(12), 758-762.

https://doi.org/10.1097/00005053-198012000-00009

McCrimmon, A.W., Smith, A.D. (2013). Review of the Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II). *Journal of Psychoeducational Assessment. 2013;31*(3):337-341. https://doi.org/10.1177%2F0734282912467756

Milo, R., & Miller, A. (2014). Revised diagnostic criteria of multiple sclerosis. *Autoimmun Rev, 13*(4-5), 518-524. https://doi.org/10.1016/j.autrev.2014.01.012

Moccia, M., Lanzillo, R., Palladino, R., Chang, K. C., Costabile, T., Russo, C., . . . Brescia Morra, V. (2016). Cognitive impairment at diagnosis predicts 10-year multiple sclerosis progression. *Mult Scler*, 22(5), 659-667. https://doi.org/10.1177/1352458515599075

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research, 151*(1), 53-79. https://doi.org/10.1016/S0377-2217(02)00578-7

Paolo, A. M., Troster, A. I., & Ryan, J. J. (1997). Test-retest stability of the California verbal learning test in older persons. *Neuropsychology, 11*(4), 613-616. https://doi.org/10.1037//0894-4105.11.4.613

PAR (2021, 28.11.21), *Brief Visuospatial Memory Test–Revised,* Obtained from https://www.parinc.com/Products/Pkey/30

Patti, F. (2009). Cognitive impairment in multiple sclerosis. *Mult Scler, 15*(1), 2-8. https://doi.org/10.1177/1352458508096684

Petersen, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *J Intern Med, 256*(3), 183-194. https://doi.org/10.1111/j.1365-2796.2004.01388.x

Rao, S. M., Leo, G. J., Haughton, V. M., St Aubin-Faubert, P., & Bernardin, L. (1989). Correlation of magnetic resonance imaging with neuropsychological testing in multiple sclerosis. *Neurology, 39*(2 Pt 1), 161-166. https://doi.org/10.1212/wnl.39.2.161

Rudick, R., Antel, J., Confavreux, C., Cutter, G., Ellison, G., Fischer, J., . . . Willoughby, E. (1997). Recommendations from the national multiple sclerosis society clinical outcomes assessment task force. Annals of Neurology, 42(3), 379-382. https://doi.org/10.1002/ana.410420318

Ruscio, J. (2008). "A probability-based measure of effect size: Robustness to base rates and other factors." *Psychological Methods 13*(1): 19-30. https://doi.org/10.1037/1082-989x.13.1.19

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychol Rev, 103*(3), 403-428. https://doi.org/10.1037/0033-295x.103.3.403

Schwid, S. R., Goodman, A. D., Weinstein, A., McDermott, M. P., Johnson, K. P., & Copaxone Study, G. (2007). Cognitive function in relapsing multiple sclerosis: minimal changes in a 10-year clinical trial. *J Neurol Sci, 255*(1-2), 57-63. https://doi.org/10.1016/j.jns.2007.01.070

Simonsen, C. S., Flemmen, H. O., Broch, L., Brunborg, C., Berg-Hansen, P., Moen, S. M., & Celius, E. G. (2021). The course of multiple sclerosis rewritten: a Norwegian population-based study on disease demographics and progression. *J Neurol, 268*(4), 1330-1341. https://doi.org/10.1007/s00415-020-10279-7

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6*(4), 317-329. https://doi.org/10.1037//1082-989x.6.4.317

Siqveland, J., Dalsbø, T. K., & Harboe, I. L. K. (2014). *Psychometric evaluation of the Norwegian version of the Wechsler Abbreviated Scale of Intelligence (WASI).* (Rapport 20). Kunnskapssenteret. https://www.fhi.no/globalassets/dokumenterfiler/rapporter/2014/rapport_2014_20_wasi.pdf

Sonder, J. M., Burggraaff, J., Knol, D. L., Polman, C. H., & Uitdehaag, B. M. (2014). Comparing long-term results of PASAT and SDMT scores in relation to neuropsychological testing in multiple sclerosis. *Mult Scler, 20*(4), 481-488. https://doi.org/10.1177/1352458513501570

Sousa, C. S., Neves, M. R., Passos, A. M., Ferreira, A., & Sa, M. J. (2018). Paced Auditory Serial Addition Test (PASAT 3.0 s): Demographically corrected norms for the Portuguese population. *Appl Neuropsychol Adult, 25*(5), 417-423. https://doi.org/10.1080/23279095.2017.1323752

Sperling, R. A., Guttmann, C. R., Hohol, M. J., Warfield, S. K., Jakab, M., Parente, M., Diamond, E. L., Daffner, K. R., Olek, M. J., Orav, E. J., Kikinis, R., Jolesz, F. A., & Weiner, H. L. (2001). Regional magnetic resonance imaging lesion burden and cognitive function in multiple sclerosis: a longitudinal study. *Arch Neurol, 58*(1), 115-121. https://doi.org/10.1001/archneur.58.1.115

Statistisk sentralbyrå. (2022, january 18th). Educational attainment of the population. Obtained from: https://www.ssb.no/en/utdanning/utdanningsniva/statistikk/befolkningens-utdanningsniva

Statistisk sentralbyrå. (2022, january 18th). Highest level of education in Oslo. Obtained from: https://www.ssb.no/en/utdanning/artikler-og-publikasjoner/highest-level-of-education-in-oslo

Strober, L. B., Christodoulou, C., Benedict, R. H., Westervelt, H. J., Melville, P., Scherl, W. F., Weinstock-Guttman, B., Rizvi, S., Goodman, A. D., & Krupp, L. B. (2012). Unemployment in multiple sclerosis: the contribution of personality and disease. *Mult Scler, 18*(5), 647-653. https://doi.org/10.1177/1352458511426735

Sumowski, J. F., Benedict, R., Enzinger, C., Filippi, M., Geurts, J. J., Hamalainen, P., Hulst, H., Inglese, M., Leavitt, V. M., Rocca, M. A., Rosti-Otajarvi, E. M., & Rao, S. (2018). Cognition in multiple sclerosis: State of the field and priorities for the future. *Neurology, 90*(6), 278-288. https://doi.org/10.1212/WNL.0000000000004977

Tintore, M., Vidal-Jordana, A., & Sastre-Garriga, J. (2019). Treatment of multiple sclerosis - success from bench to bedside. *Nat Rev Neurol, 15*(1), 53-58. https://doi.org/10.1038/s41582-018-0082-z

Tombaugh, T. N. (2006). A comprehensive review of the Paced Auditory Serial Addition Test (PASAT). *Arch Clin Neuropsychol, 21*(1), 53-76. https://doi.org/10.1016/j.acn.2005.07.006

Van Schependom, J., D'Hooghe M, B., Cleynhens, K., D'Hooge, M., Haelewyck, M. C., De Keyser, J., & Nagels, G. (2015). Reduced information processing speed as primum movens for cognitive decline in MS. *Mult Scler, 21*(1), 83-91. https://doi.org/10.1177/1352458514537012

Vanotti, S., Smerbeck, A., Benedict R.H. B. & Caceres, F. (2016). A new assessment tool for patients with multiple sclerosis from Spanish-speaking countries: validation of the Brief International Cognitive Assessment for MS (BICAMS) in Argentina, *The Clinical Neuropsychologist*, *30:7*, 1023-1031, https://doi.org/10.1080/13854046.2016.1184317

Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analyses of age-cognition relations in adulthood: estimates of linear and nonlinear age effects and structural models. *Psychol Bull, 122*(3), 231-249. https://doi.org/10.1037/0033-2909.122.3.231

Woods, S. P., Delis, D. C., Scott, J. C., Kramer, J. H., & Holdnack, J. A. (2006). The California Verbal Learning Test--second edition: test-retest reliability, practice effects, and reliable change indices for the standard and alternate forms. *Arch Clin Neuropsychol, 21*(5), 413-420. https://doi.org/10.1016/j.acn.2006.06.002

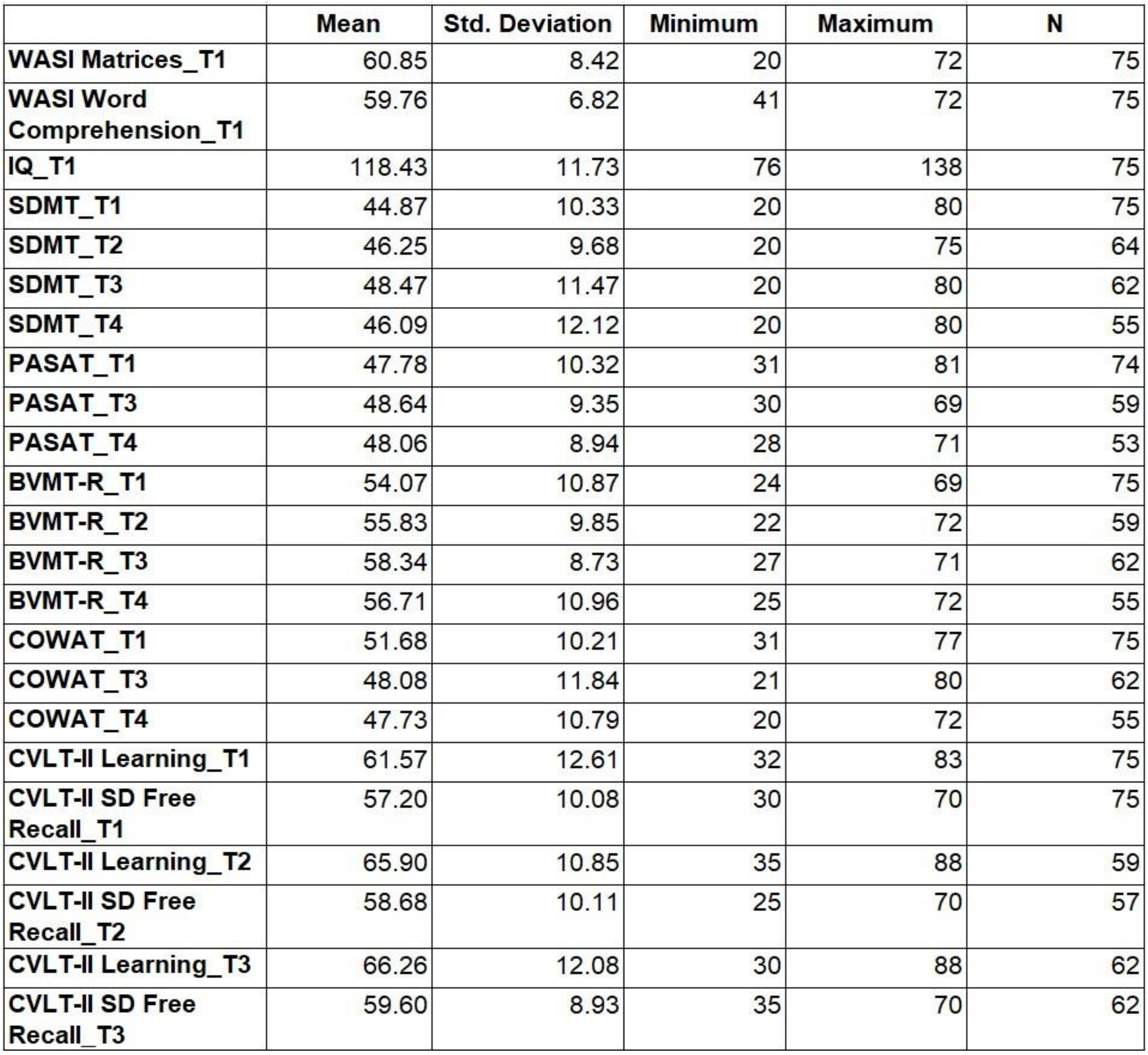
Zgaljardic, D. J., & Benedict, R. H. (2001). Evaluation of practice effects in language and spatial processing test performance. *Appl Neuropsychol, 8*(4), 218-223. https://doi.org/10.1207/S15324826AN0804_4

Zimprich, D., & Martin, M. (2002). Can longitudinal changes in processing speed explain longitudinal age changes in fluid intelligence? *Psychol Aging, 17*(4), 690-695. https://doi.org/10.1037/0882-7974.17.4.690

Appendix A: Descriptive statistics of standardized neuropsychological test scores at all four time points and histograms of neuropsychological test scores at T1 and T4

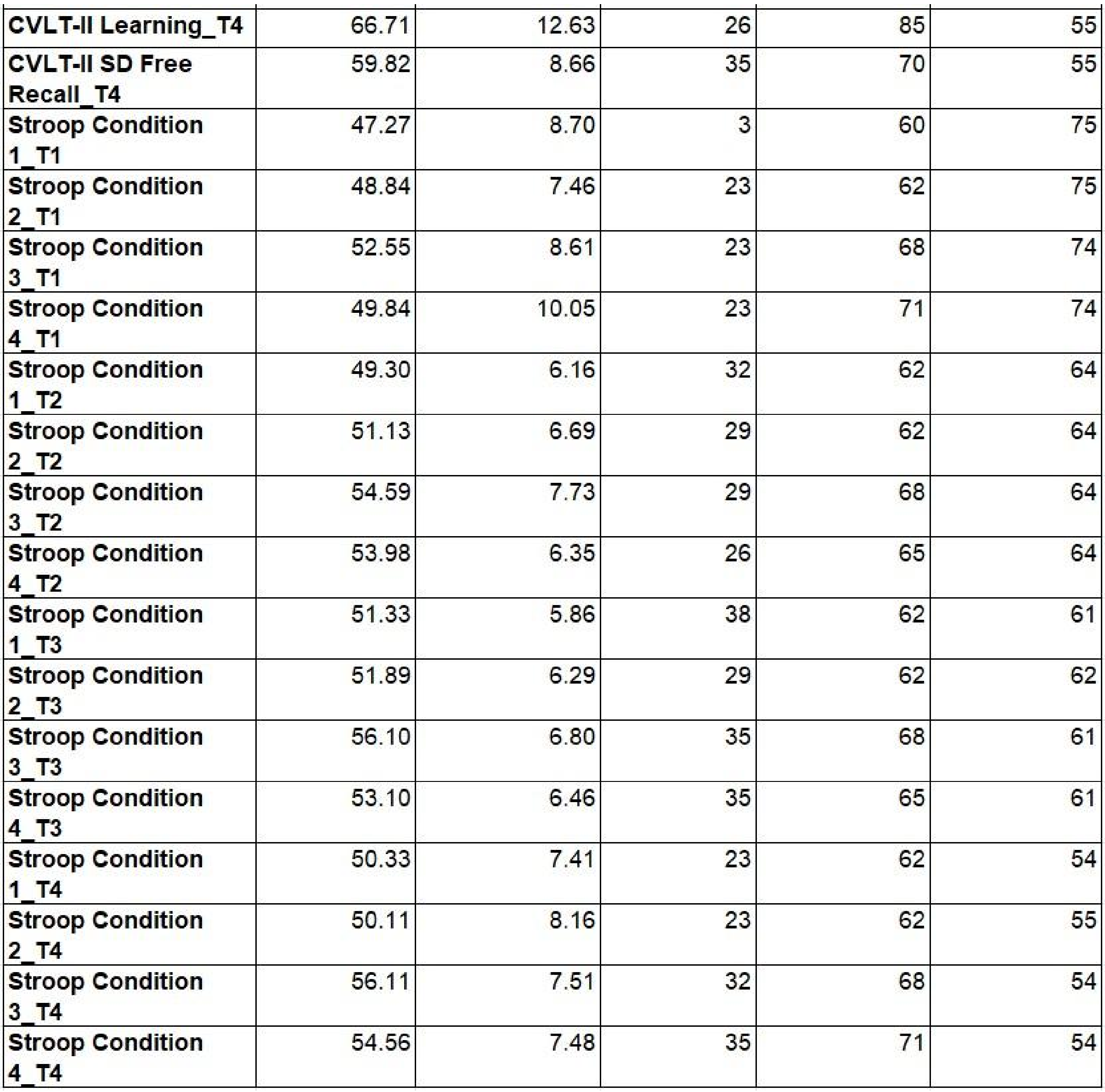
Table A1

Descriptive statistics of all the neuropsychological tests at all four test points. The scores are standardized and given in T-scores, with the exception of IQ.



	Mean	Std. Deviation	Minimum	Maximum	N
WASI Matrices_T1	60.85	8.42	20	72	75
WASI Word Comprehension_T1	59.76	6.82	41	72	75
IQ_T1	118.43	11.73	76	138	75
SDMT_T1	44.87	10.33	20	80	75
SDMT_T2	46.25	9.68	20	75	64
SDMT_T3	48.47	11.47	20	80	62
SDMT_T4	46.09	12.12	20	80	55
PASAT_T1	47.78	10.32	31	81	74
PASAT_T3	48.64	9.35	30	69	59
PASAT_T4	48.06	8.94	28	71	53
BVMT-R_T1	54.07	10.87	24	69	75
BVMT-R_T2	55.83	9.85	22	72	59
BVMT-R_T3	58.34	8.73	27	71	62
BVMT-R_T4	56.71	10.96	25	72	55
COWAT_T1	51.68	10.21	31	77	75
COWAT_T3	48.08	11.84	21	80	62
COWAT_T4	47.73	10.79	20	72	55
CVLT-II Learning_T1	61.57	12.61	32	83	75
CVLT-II SD Free Recall_T1	57.20	10.08	30	70	75
CVLT-II Learning_T2	65.90	10.85	35	88	59
CVLT-II SD Free Recall_T2	58.68	10.11	25	70	57
CVLT-II Learning_T3	66.26	12.08	30	88	62
CVLT-II SD Free Recall_T3	59.60	8.93	35	70	62

Table A1 (continued)



CVLT-II Learning_T4	66.71	12.63	26	85	55
CVLT-II SD Free Recall_T4	59.82	8.66	35	70	55
Stroop Condition 1_T1	47.27	8.70	3	60	75
Stroop Condition 2_T1	48.84	7.46	23	62	75
Stroop Condition 3_T1	52.55	8.61	23	68	74
Stroop Condition 4_T1	49.84	10.05	23	71	74
Stroop Condition 1_T2	49.30	6.16	32	62	64
Stroop Condition 2_T2	51.13	6.69	29	62	64
Stroop Condition 3_T2	54.59	7.73	29	68	64
Stroop Condition 4_T2	53.98	6.35	26	65	64
Stroop Condition 1_T3	51.33	5.86	38	62	61
Stroop Condition 2_T3	51.89	6.29	29	62	62
Stroop Condition 3_T3	56.10	6.80	35	68	61
Stroop Condition 4_T3	53.10	6.46	35	65	61
Stroop Condition 1_T4	50.33	7.41	23	62	54
Stroop Condition 2_T4	50.11	8.16	23	62	55
Stroop Condition 3_T4	56.11	7.51	32	68	54
Stroop Condition 4_T4	54.56	7.48	35	71	54

The figures below are generally organized in a way that the left figure shows a histogram of a test score or other variable at T1, and the right figure shows a histogram of the same test or variable at T4. Some histograms at T4 have a bigger variety of possible T-scores as a consequence of imputation (e.g. CVLT-II long-term memory was standardized according to Z-scores, which result in converted T-score interval jumps of five. Generated scores due to imputation do not follow this same five T-score interval jump, meaning a participant can get an imputed T-score of 53 instead of either 50 or 55, where the latter would be the result of a direct conversion from Z-scores as done for T1).

Figure A1

Histogram of SDMT T-score distribution at T1

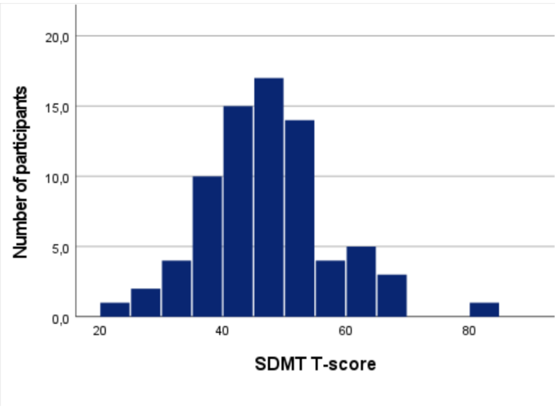


Figure A2

Histogram of SDMT T-score distribution at T4

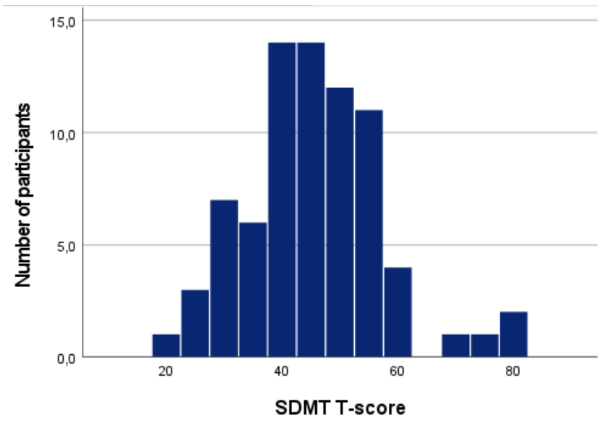


Figure A3

Histogram of PASAT T-score distribution at T1

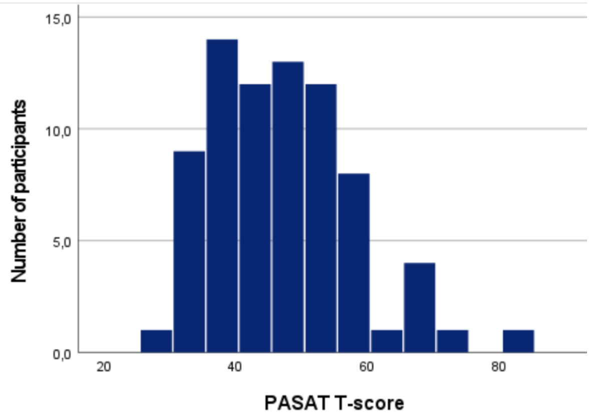


Figure A4

Histogram of PASAT T-score distribution at T4

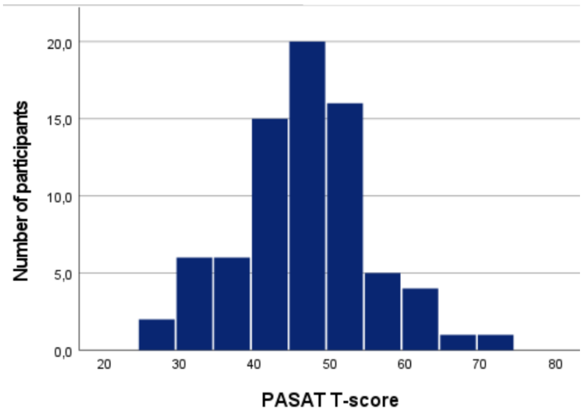


Figure A5

Histogram of BVMT-R T-score distribution at T1

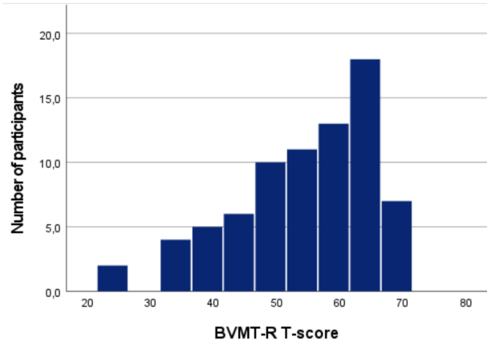


Figure A6

Histogram of BVMT-R T-score distribution at T4

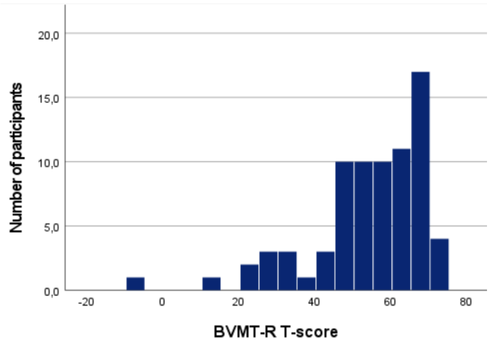


Figure A7

Histogram of COWAT T-score distribution at T1

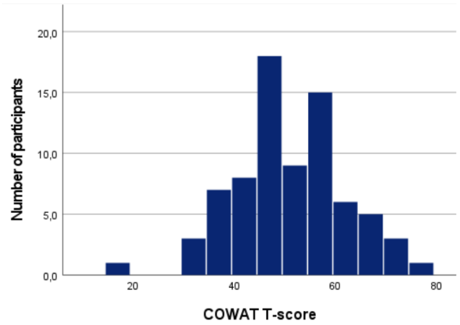


Figure A8

Histogram of COWAT T-score distribution at T4

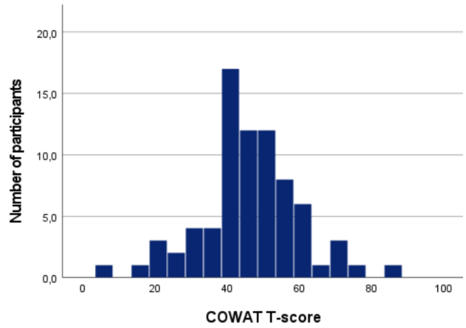


Figure A9

Histogram of CVLT-II learning T-score distribution at T1

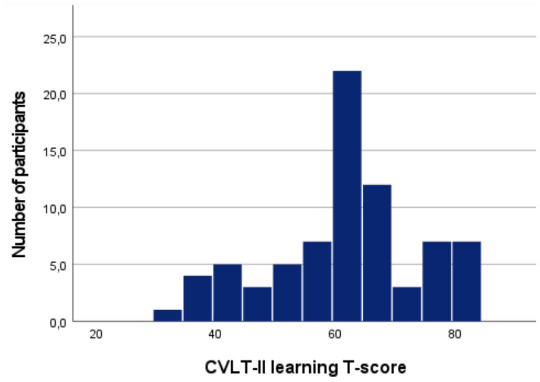


Figure A10

Histogram of CVLT-II learning T-score distribution at T4

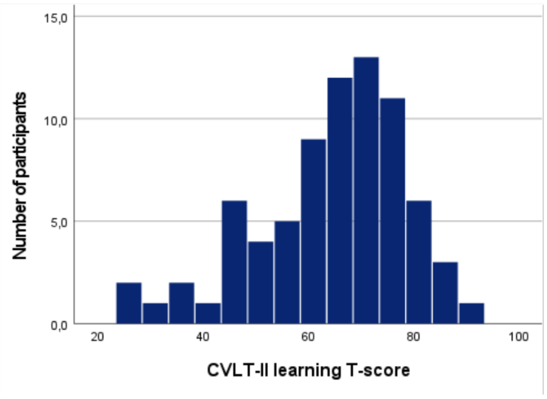


Figure A11

Histogram of CVLT-II long-term learning T-score distribution at T1

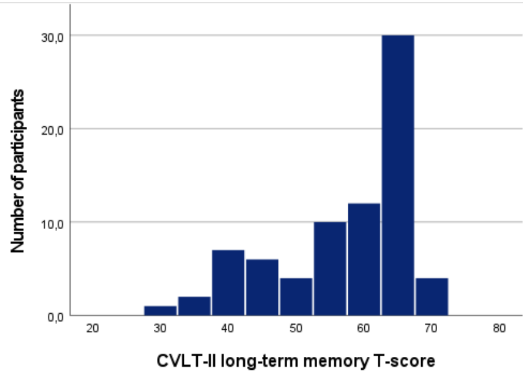


Figure A12

Histogram of CVLT-II long-term learning T-score distribution at T4

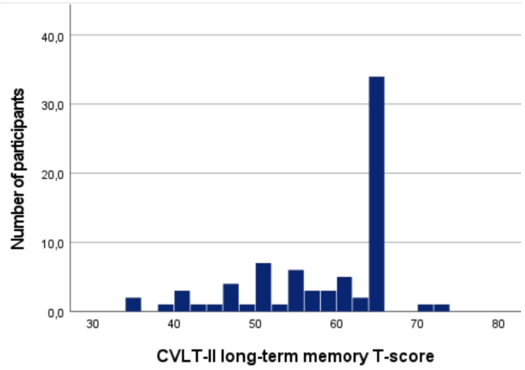


Figure A13

Histogram of D-KEFS Stroop condition 3 T-score distribution at T1

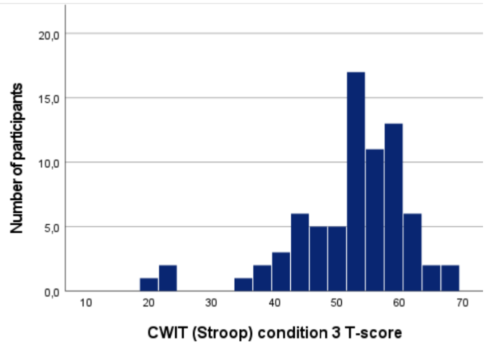


Figure A14

Histogram of D-KEFS Stroop condition 3 T-score distribution at T4

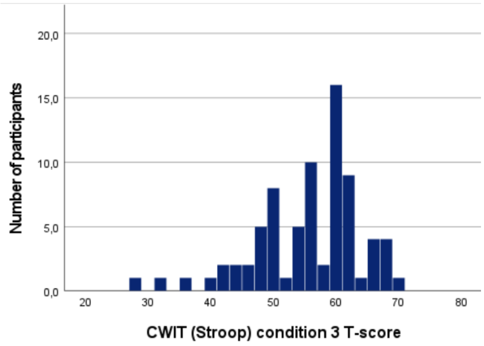


Figure A15

Histogram of D-KEFS Stroop condition 4 T-score distribution at T1

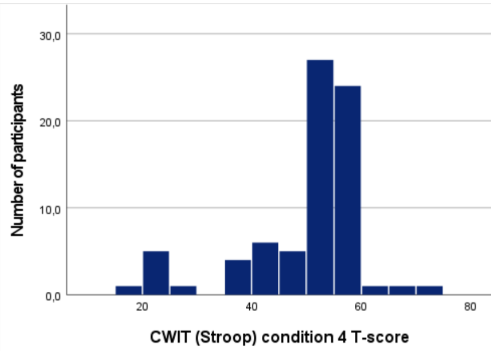


Figure A16

Histogram of D-KEFS Stroop condition 4 T-score distribution at T4

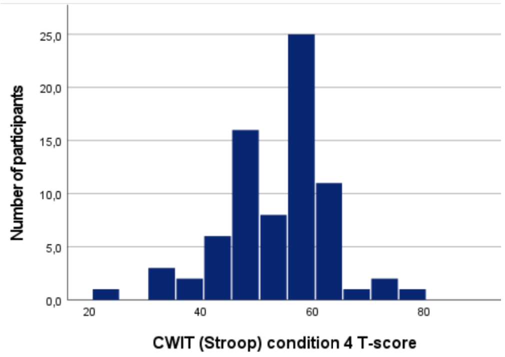


Figure A19

Histogram of GCF T-score distribution at T1

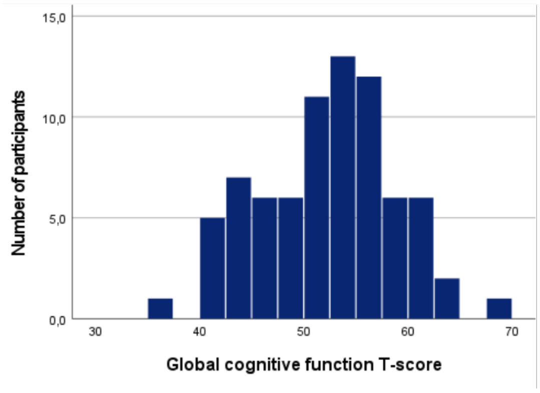


Figure A20

Histogram of GCF T-score distribution at T4

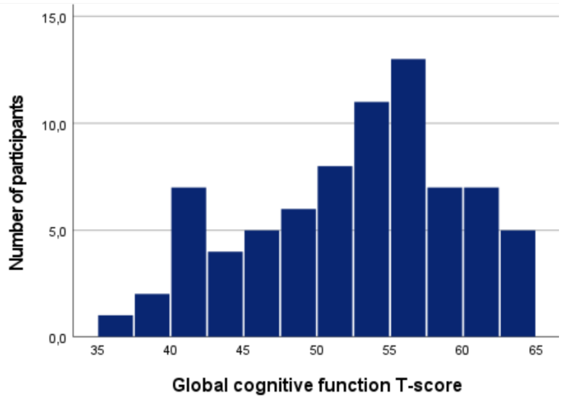
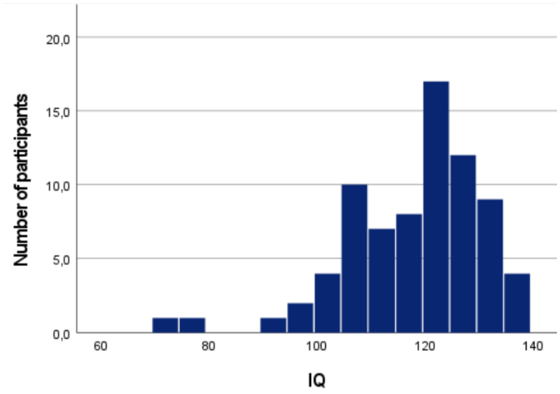


Figure A21

Histogram of IQ distribution at T1 calculated by WASI word comprehension and matrices



Appendix B: Pie charts of participants current and previous treatment

Figure A22

Pie chart of participants' current treatment

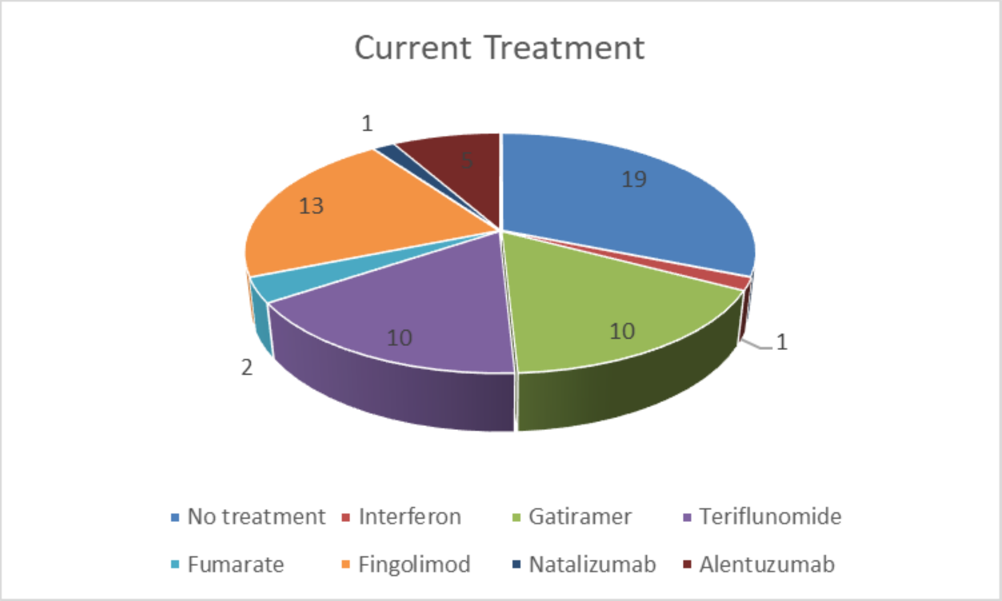
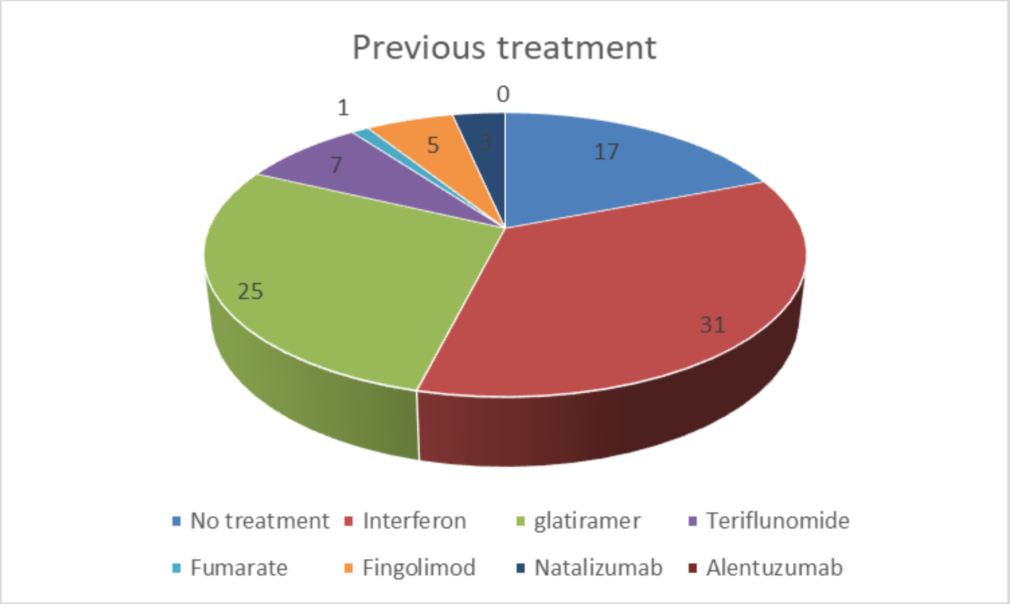


Figure A23

Pie chart of participants' previous treatment


Previous treatment
■ No treatment
■ Interferon
■ glatiramer
■ Teriflunomide
■ Fumarate
■ Fingolimod
■ Natalizumab
■ Alentuzumab