Utrecht University

Efficient neuropsychological assessment: Reducing the duration of test research

Reviewing and developing abbreviated neuropsychological tests and evaluating their psychometric and diagnostic quality

Master Thesis Neuropsychology

Author: Rianne L. Penninga

Email: r.l.penninga@students.uu.nl

Student number: 3520595

Date: July 10, 2015

Supervisors:

dr. M. J. E. van Zandvoort (m.vanzandvoort@uu.nl)

I. M. C. Huenges Wajer MSc (i.m.c.huengeswajer@umcutrecht.nl)

Table of Contents

Abstract

Current changes in health and health centers require efficient neuropsychological assessment. This thesis evaluates whether reduction in assessment time of neuropsychological evaluation can be accomplished by using abbreviated test versions. Abbreviated forms of frequently used neuropsychological tests were selected from literature or developed. Multiple methods were used to examine not only their psychometric quality, but also to compare their diagnostic quality with the original tests. The results showed that some abbreviated forms can replace original tests in neuropsychological evaluation for diagnostic purposes. However, other abbreviated forms should not even be used as screening instruments. Although the use of abbreviated forms can definitely contribute to efficient neuropsychological assessment, it is important that the quality of these forms is examined carefully.

1. Introduction

In recent years, the neuropsychological examination (NPE) of the cognition of patients has become more common in health centers (Lezak, 2012). By observing behavior and interpreting test results, neuropsychologists are able to draw conclusions about cognitive (dys)functioning, e.g. in order to answer diagnostic questions. Since medical attention and treatment is improving and the global population is ageing, it is expected that the frequency in which NPE is used will increase rapidly (Deelman, 2009). Simultaneously, the current financial climate in health centers requires cost-effective assessment. Consequently, clinicians feel the necessity for efficient NPE, as can be accomplished by minimizing its duration.

It is beneficial for patients to reduce the duration of NPE as well. NPE is not only time consuming for them too but moreover it can cause distress and frustration, for example if patients are aware of the decline in cognitive performance (Fastenau, Denburg & Mauer, 1998). In addition, many patients that are seen by neuropsychologists often experience fatigue, for example after stroke (Ingles, Eskes & Philips, 1999), traumatic brain injury (Zino & Ponsford, 2006), Parkinson's Disease (Karlsen, Larsen, Tandberg & Jørgensen, 1999), or depression (Kinsinger, Lattie & Mohr, 2010). Fatigue can even result in patients being too tired to participate in various time consuming neuropsychological (NP) tests. Besides that distress, frustration or fatigue is a burden for patients, test results can be easily misinterpreted when they are influenced by these factors (Hendriks, Kessels, Gorissen & Schmand, 2010). In conclusion, from different points of view it is very beneficial to minimize the duration of NPE.

It is therefore not surprising that in the last few years an increasing number of studies have considered the use and the quality of screening instruments for brief mental examination. Nowadays, many different screening tests are used in clinical settings (Cullen et al., 2007; Uttner et al., 2013). However, review articles state that the diagnostic accuracy of screening instruments is often inconsistent (Cullen et al., 2007), i.e. in dementia-screening tests (Appels & Scherder, 2010). Moreover, according to Lezak and colleagues (2012), diagnosing always requires more than screening and the use of screening tests is often even inappropriate to use in NPE for diagnostic purposes. So, clinicians cannot rely on screening tests when answering diagnostic questions.

Therefore, this thesis focuses on other possibilities to minimize the duration of NPE in which the diagnostic quality is maintained. One possibility is to focus on the selection of the minimal number of NP tests that is necessary in NPE. Unfortunately, it is impossible to select one set of NP tests which is appropriate for all diagnostic questions, since the selection of

tests depends on different aspects, including the goal of examination (Deelman et al., 2009), the psychometric quality of the tests, and practical issues (Lezak et al., 2012). Consequently, no general ultimate test battery can be defined.

Another way to reduce the duration of NPE is to use abbreviated versions of tests. For some tests, different abbreviated test versions are either published or used in clinical settings. For other tests, abbreviated test versions are not yet developed or abbreviation is not possible without changing the test itself. The goal of this thesis is to evaluate the quality of abbreviated test versions that are either developed or that will be developed in this thesis. Since not only the time gain but also the psychometric and diagnostic quality of short forms is important for efficient neuropsychological assessment, these aspects will be taken into account when making an educated judgment about the use of abbreviated forms as part as efficient NPE.

2. Methods

This study consists of three parts. Firstly, a data-driven approach is used to select NP tests that are frequently used for diagnostic purposes. Secondly, for every selected test the literature about abbreviated test versions is reviewed (theory-driven). Thirdly, when abbreviation of a selected NP test was possible, promising or newly developed short versions were compared with the original tests (data-driven) in terms of reliability, their relation with each other and with third variables, consistency between the ranking positions of the outcomes and diagnostic agreement.

2.1 Selecting frequently used tests

As discussed in the introduction, it is not possible to select one set of NP tests that is appropriate to answer all diagnostic questions. Therefore, tests were selected that are frequently seen as appropriate by neuropsychologists when answering diagnostic questions related to a possible neurologic defect. It is of course more beneficial to develop optimal short forms of NP tests that are frequently used than tests that are used less often, so that time savings can occur more frequently as well. For this thesis, the ten most frequently used NP tests are selected for further evaluation about possible abbreviation.

For this selection, NPE is considered for which no standard test battery is used but for which neuropsychologists have selected NP tests that they consider, based on their expertise, as relevant when answering diagnostic questions. Hereby, it can be assumed that the most frequently used tests that were selected for further evaluation are often considered as useful when answering diagnostic questions. Another requirement for the selection of frequently used tests is that the tests are standardized and worldwide both respected and used by neuropsychologists. In this way, the results of this thesis can be used in a broad community of neuropsychologists worldwide.

A register of the use of NP tests that meets both described requirements was found in the neurology department University Medical Center of Utrecht (UMCU). It is therefore assumed that the frequently used NP tests revealed by this register are NP tests that are frequently chosen when answering diagnostic questions worldwide. The dataset of the UMCU contains data from 451 outpatients (198 female) who were seen at the neurological department by a neuropsychologist for various diagnostic purposes (patients participating in the HAMLET-study are excluded for analysis), between October 2003 and July 2013. The native language of all patients included in the dataset is Dutch, the mean age of the patients

was 52.1 years (SD = 16.6) at the time of assessment. The most frequently used tests revealed by this register can be found in table 1 (see paragraph 3).

2.2 Literature review

For every selected test, it is examined if there are any studies published about the reduction of the duration of this test. These studies were identified by searching databases PubMed, Scopus and Google Scholar, using the key terms 'duration', 'abbreviated', 'short' and 'parallel version' in combination with the name of the test. When a short version of a test was found, it's quality was described. If possible, the current study elaborated on previous research that was revealed by the literature review.

2.3 Comparing the short versions with the original tests

The third and last part of this study consisted of the comparison of the short version(s) with the original test in terms of reliability, their relation with each other and with third variables (equivalence), consistency between the ranking positions of the outcomes and diagnostic agreement. More information about the methods that are used when comparing the test versions is written in the next paragraphs. For all analyzes, the statistical power will be reported. However, the statistical power can only be defined when the data are normally distributed (Field, 2009). Note that a summary of the results of all analyzes can be found in appendix 1.

For the comparison between the short versions and the original tests, data of all separate test items needed to be available. Therefore, data was collected and inserted into SPSS (version 22.0).

Participants

To collect this data, acquaintances of researchers at the department were approached. Informed consent was obtained prior to the assessment, no financial compensation was provided. The NP tests were administered as part of a larger test battery. The same instructions and feedback rules were applied to all participants. Based on a small survey, participants that once had a cerebrovascular insult (stroke or TIA), head injury with loss of consciousness for at least five minutes or chronic substance abuse were excluded from the sample.

Data of 21 healthy adults (9 female) is collected. The participants had a mean age of 56.9 years (SD = 12.8). The levels of education were classified according to Verhage (1964), the mean level of education was 5.7 (SD = 1.4).

Reliability

Reliability is one of the most important qualities when evaluating the adequacy of any type of psychological test (Goodwin, 2008). A measure of behavior should be consistent to be informative, and reliability, the degree in which a measure is consistent, is therefore one of the five criteria applied by the Dutch psychological association (COTAN; Egberink, Janssen & Vermeulen, 2014). It is therefore of great importance that not only original NP tests but also the abbreviated test versions have a good reliability. The inter-item correlation, the most often used measure of reliability which is expressed by Cronbach's α, is used to examine the reliability in both the original as the short test versions. SPSS (version 22.0) is used to calculate these inter-item correlations. Alpha coefficients were interpreted as: $\alpha \geq 0.9$: 'very high', $0.9 > \alpha \geq 0.8$: 'high', $0.8 > \alpha \geq 0.7$: 'adequate', $0.7 > \alpha \geq 0.6$: 'marginal', and $\alpha < 0.6$: 'low' (Strauss, Sherman & Spreen, 2006). Note that although these interpretations are common when interpreting the internal consistency of NP tests, more conservative interpretations are published as well (Nunally & Bernstein, 1994). Ideally, the reliability of the short version is just as good as the reliability of the original test. In this study, a test version with a reliability that is lower than adequate is not approved for the use as a stand-alone measure.

Equivalence

Since the goal of this study was to select short versions that can replace the original tests in NPE used for diagnostic purposes, it is of great importance that the short versions measure the same construct as the original tests. Three methods are used to evaluate the equivalence. For some tests, it was suitable to use a fourth measure. If this was the case, explanatory notes can be found in the paragraph of the specific test (e.g. paragraph 4.2.2).

Firstly, if the short version measures the same construct as the original test, the scores obtained from both versions should relate closely to each other. Therefore, the correlation coefficients between the scores obtained from the two different test versions are calculated. If the short form is representative to the original form, the relation should closely approximate 1: a perfect relation. Correlation coefficients were interpreted as follows: $r \geq 0.7$: 'strong relationship', $0.7 > \alpha \geq 0.5$: 'moderate relationship', or $0.5 > \alpha \geq 0.3$: 'weak relationship (Field, 2009). When more than one short version per test is evaluated, all correlation coefficients between the short versions and the original test were compared to each other, to see if all short versions related to the original test in the same way. To see if there was a significance difference between these correlations, the procedure that is described in the last section of this paragraph was applied.

Secondly, when the short version measures the same construct as the original test, the relation between the original test and demographic variables should be in the same

range as the relation between the short version and these demographic variables. This is also called parallelism (Fastenau, Denburg & Mauer, 1998) and is examined by comparing the correlation coefficients of the different test versions with the demographic variables age, sex and level of education. To test if the relations were indeed in the same range, the procedure described in the last section of this paragraph was applied.

Thirdly, if the short version measures the same construct as the original test, the relation between the original test and an external construct should be in the same range as the relation between the short version and this external construct. For every selected NP test an outcome of a different NP test is chosen as external construct. When possible, external constructs are chosen that measure the same cognitive domain. However, of importance for this study is whether the relations of the short and original test with the external construct are in the same range and not whether there is a significant relation with this external construct. If the short version is equivalent to the original version, there should not be any difference between the relations of these test versions with the external construct. To test if the relations were indeed in the same range, the procedure described in the last section of this paragraph was applied.

For the analyzes, correlation coefficients are calculated with the use of SPSS, version 22.0. Pearson's correlation coefficient is used when data are normally distributed, and Spearman's correlation coefficient is used when data are non-normally distributed.

For all three measures of equivalence, correlations needed to be compared to each other to see if they are in the same range or differ significantly from each other. Therefore, each correlation was first converted into a z-score using Fisher's r to z transformation. Then, Steiger's equations 3 and 10 (1980) were used to compute the asymptotic covariance of these measures. At last, the estimates were used in an asymptotic z-test to see if the correlation coefficients did not differ significantly from each other, i.e. if the relations were in the same range. Two-tailed tests were used, because it is unclear in advance which test versions has the strongest relation with another variable. During these analyzes, three details were taken into account. This is illustrated with the following example. When it was estimated whether a short form has the same relation with age as the original test, it was taken into account that these two correlations are dependent, since age is part of both correlations. Furthermore, it was taken into account that the relations are found in the same sample. Lastly, there was corrected for the correlation between the short and the original test version. All the formulas that were necessary for these analyzes have been programmed in online software that is used when estimating whether there is a difference between the correlations (Lee & Preacher, 2013).

Consistency between scores

When a short test version measures the same qualities as the original test, participants should score equally (bad or good) on the short version as on the original test. To determine if this was the case, the ranking positions for the scores on the different test versions were computed. Then, the ranking positions of a short form and the original test were compared to each other with a Wilcoxon Signed Rank Test. There is chosen for this test because data do not need to be normally distributed and because the samples can be related. In addition, it is more appropriate to use a Wilcoxon Signed Rank Test than a t-test in a sample that is smaller than 25 people, as was the case in this research. Analyzes were conducted using SPSS (version 22.0). When the difference between the ranking positions were not significant, it was assumed that there is consistency between scores: Participants obtained the same rank in the sample when different test versions were used.

Diagnostic agreement

Because this thesis focuses on short forms that should be able to replace original tests in NPE used for diagnostic purposes, it is of great importance to evaluate whether the short form and original test reach the same conclusion about the performance of a participant. To decide whether the short form and the original form point to the same diagnostic category, the classification of the scores obtained from the different test versions were compared. The categories that were used are 'impaired', 'below average', 'average', 'above average' and 'excellent', and these categories were assigned to the scores according to Lezak and colleagues (2012): SD < 2.0: 'impaired', -1.0 > SD ≥ -2.0: 'below average', 1.0 > SD ≥ -1.0: 'average', 2.0 > SD ≥ 1.0: 'above average', and SD ≥ 2.0: 'excellent'. Common used normative data were used to interpret the scores. More details about this normative data can be found in the method sections of each separate test. If normative data for the short forms was not available, the scores were linearly transferred so that they could be compared with the normative data that was available for the original test. The exact formulas for these transformations are discussed in the method sections of each separate test as well.

Note that for some tests, there was reason to believe that the scores on the short forms could not be linearly transformed into the scores for original tests. It went beyond the scope of this thesis to calculate non-linear formula's to transform this data. Therefore, when data could not be linearly transformed and when there was no normative data available for the short forms itself, diagnostic agreement was not evaluated for these tests versions. If this was the case, explanatory notes about the non-linearity can be found in the method sections of the evaluated tests.

In previous studies about the development and evaluation of short forms, diagnostic agreement is often not assessed. When diagnostic agreement is taken into account, the

percentage of agreement is often used as a measure of diagnostic agreement. This measure does however not correct for chance. Since the diagnostic categories should not overlap by chance but should overlap because there actually is diagnostic agreement, the author decided to use a different measure, namely Cohen's Kappa. This measure takes the percentage of agreement into account but corrects it for chance. The values of Cohen's Kappa were interpreted as: $\kappa \geq 0.81$: 'almost perfect', $0.81 > \kappa \geq 0.61$: 'substantial', $0.61 > \kappa \geq 0.41$: 'moderate', $0.41 > \kappa \geq 0.21$: 'fair', and $\kappa < 0.21$: 'poor' (Landis & Koch, 1977). Diagnostic agreement between a short form and the original test should be almost perfect, otherwise the short form cannot be used as a replacement of the original test in NPE.

3. Results: The selected tests for further evaluation

To determine which NP tests are most often used when answering diagnostic questions, a dataset of the neurology department of the UMCU is used (see paragraph 2.1, selecting frequently used tests). The ten most frequently administered NP tests can be found in table 1. For these tests, the possibilities and implications of the use of abbreviated test versions for diagnostic purposes will be evaluated in this thesis. More information about these tests can be found in the ten next paragraphs that are about each test separately.

Since it is known that experts of the neuropsychology have the impression that especially the Visual Elevator (subtest of the Test of Everyday Attention) is more time consuming than necessary, the abbreviation of this test will also be evaluated in this thesis (see paragraph 4.11). This subtest is administered in 38.8 % of the patients of whom the data is analyzed and has a ranking position of 21.

In sum, eleven tests are selected for further examination and will be discussed in the next eleven paragraphs.

Table 1. *The ten most frequently used NP tests at the neurology department of the UMCU, their corresponding ranking position and the percentage of the patients in which the test was administered.*

Test	Rank	Percentage
WAIS Digit Span	1	94.5
Rey Auditory Verbal Learning Test	2	93.6
Verbal Fluency, semantic (animals)	3	93.1
Boston Naming Test	4	91.4
Verbal Fluency, phonological (Letter N and Letter A)	5	91.1
Trail Making Test	6	82.0
Rey-Osterrieth Complex Figure (copy & delayed recall)	7	81.4
Brixton Spatial Anticipation Test	8	72.7
Judgment of Line Orientation	9	71.4
Token Test	10	64.1

4. Results: Evaluation of each selected test

4.1 WAIS Digit Span

Literature Review

The Digit Span is one of the subtests of the third and fourth edition of the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 2008) and it is the second most used neuropsychological instrument for attention (Rabin, Barr & Burton, 2005). The Digit Span Forward is the first part of the subtest, in which patients have to remember and repeat a string of numbers. In the second part of the WAIS Digit Span (Digit Span Backward), strings of numbers have to be recited in reversed order. In both parts, the first string consists of two digits. After every second string an additional digit is added, until the string consists of nine numbers (Digit Span Forward) or eight number (Digit Span Backward). When responses to two strings with the same number of digits are incorrect (i.e. when two successive failures are made) the administration is discontinued. For this reason, the administration time of the WAIS Digit Span depends on the errors that are made.

During the literature search about the abbreviation of the WAIS Digit Span, 15 articles are selected for further examination based on their titles and abstracts. Although in several studies items of various subtests of the WAIS are eliminated to create an abbreviated intelligence test, the Digit Span subtest is consistently left unchanged (e.g. Wymer, Rayls, & Wagner, 2003; Meyers, Zellinger, Kockler, Wagner & Miller, 2013). No studies about the abbreviation of the Digit Span subtest itself are found.

There are several options to reduce the administration time of the WAIS Digit Span. One option is to stop the administration after one failure, instead of two successive failures. Blackburn and Benton (1957) modified the WAIS III Digit Span: The administration was stopped after three successive failures instead of two. They found that this modified version of the WAIS III Digit Span had higher test-retest reliability than the original version. Hence, it does not seem beneficial for the quality of the test to stop the administration after only one failure.

Stopping a response after a certain amount of time could also reduce the administration time. Babikian, Bonne, Lu and Arnold (2006) investigated the effect of adding a time variable to the WAIS Digit Span. They found a strong relation between the time variable and the accurate detection of feigned cognitive symptoms. This finding is very interesting and may be useful in certain clinical purposes. However, the goal of this study is not to modify NP tests so that they detect feigned symptoms but to investigate the

possibilities to use short test versions that measure the same construct as the original test version. The use of a time variable to reduce the WAIS Digit Span will therefore not be examined in this thesis.

Summarizing, the literature review did not reveal any short version of the WAIS Digit Span. There are two possible ways in which the administration time of the test could be reduced, but, as discussed, these ways do not seem suitable. In addition, the administration time of the WAIS Digit Span is already restricted, since the administration is discontinued when responses to two strings with the same number of digits are incorrect. For these reasons, no short forms of the WAIS Digit Span will be either evaluated or developed in this study.

4.2 Rey Auditory Verbal Learning Test

4.2.1 Literature Review and Introduction

The literature search revealed eleven articles that were selected for further examination based on their titles and abstracts.

The Verbal Learning Test is initially developed by Rey (1964) and measures the declarative memory, like other neuropsychological tests (e.g. the California Verbal Learning Test; Delis, Kramer, Kaplan & Ober, 1987). More specifically, the test obtains measures of immediate memory span, new learning, retroactive and proactive interference, and recognition (Strauss et al., 2006). In the Verbal Learning Test, fifteen monosyllabic words are presented and repeated in five subsequent trials. After each trial, patients have to repeat all the words they remember (free recall). After a delay of approximately fifteen minutes, there is not only an additional free recall trial but also a recognition trial, in which participants have to decide whether the presented words were also in the original trial. The words can be presented both verbal as visual, which effects scores on the test differently (van der Elst, van Boxtel, van Breukelen & Jolles, 2005). In the current study, the Rey Auditory Verbal Learning Test (RAVLT) will be considered, since the verbal mode of presentation is the most common. Different alternate forms for the original fifteen-word RAVLT are developed, so that learning effects during a follow-up can be minimized (see inter alia Geffen, Butterworth & Geffen, 1994 and Uchiyama et al., 1995). A Dutch version of the RAVLT was developed by Brand and Jolles (1985).

Apart from parallel forms, several short forms of the RAVLT are developed. These forms consist of five trials as well, but include less than fifteen words. For example, an eight-word version can be found in the *Amsterdamse Dementie-Screeningtest* (Amsterdam Dementia Screening Test; Lindeboom & Jonker, 1989). Also, the ten first words of the original RAVLT are often used as a short version (Lezak et al., 2012). There is of course less administration time needed for these short versions. However, these short versions are often too easy for both controls and patients. A ceiling effect is often present and the short versions are therefore not useful in the largest part of the (patient) population (Leak et al, 2012). For this reason, the reduction of the administration time of the RAVLT by presenting less than fifteen words will not be further evaluated in this thesis.

Another option to reduce the administration time of the RAVLT is to reduce the number of trials, which is originally five. In a study of van der Elst and collegues (2005), in which normative data for more than 1800 healthy participants is presented, the total score of the free recall is not only obtained for trials 1 to 5, but also for only the trials 1 to 3.

Interactions with several demographic variables are displayed for both total scores, but the specific differences between these scores is not discussed. The three-trial version of the RAVLT is also been used in a study about the influence of several specific aspects of cancer on cognition (Schilder, van Dijk, Meinhardt, van Dam & Schagen, 2001). However, the difference between the three-trial version and the original five-trial version is not discussed. Therefore, the current study focuses on the possibilities and implications of the use of the three-trial version instead of the original five-trial version of the RAVLT.

4.2.2 Methods

The short form that will be evaluated in this thesis consists of trial 1 to 3 of the RAVLT. This version will be compared to the original test with 5 trials. Information about the participants and about the methods that are used when comparing the short version with the original test can be found in paragraph 2.3.

For this test, the inter-item reliability will not be analyzed because the total score on the trials of the RAVLT and not the words (that can be viewed as 15 items) are of interest. That is, the number of correctly recalled words is used to compute a score for the RAVLT and, in general, it is not taken into account which words are recalled. Therefore, the inter-item reliability is not used when comparing the representativeness of the short version.

The measures of equivalence will be analyzed as described in paragraph 2.3 (the general methods). For the comparison of the relations of the two test versions with an external construct, the relations with the score on the delayed recall trial of the Rey-Osterrieth Complex Figure Test are considered. For this test, a fourth index of equivalence will be computed. That is, it will be evaluated if the two versions can both equally predict the free recall score of the RAVLT.

The consistency between the ranking positions of the participants for the two test versions will be analyzed as described in the general methods (paragraph 2.3). The diagnostic agreement between the two test versions will however not be computed, since it is assumed that there is no linear relationship between the version with three trials and the original five-trial version. Particularly, the study of van der Elst and collegues (2005) shows that the difference between the trials does not occur in a linear fashion: The learning effect declines for each subsequent trial.

4.2.3 Results

Equivalence

To determine if the total score derived from the short form is representative to the total score derived from the original form, the relation between these two total scores is examined. Spearman's correlation coefficient is used to define this relation, because the data are not normally distributed. The correlation is strong (r_s = 0.97, p < 0.01), so the scores of the short form are representative to the scores of the original form.

As a second index of equivalence, the relations between the original version and demographic variables (namely age, sex and education) are compared to the relations between the abbreviated form and these same demographic variables. It is expected that the abbreviated form relates to these variables in the same way (parallelism). Firstly, the Spearman's correlation coefficients are computed. These show that only age is significantly correlated to the immediate recall of words in the RAVLT, whereby the performance of younger participants is better than that of older participants (r_s = -0.57, p > 0.01 for the original test; r_s = -0.60, p > 0.01 for the short form). The correlations between the total scores and respectively sex and level of education are not significant in both forms of the RAVLT (see table 2). Secondly, the z-scores for the difference between the correlations are computed (see paragraph 2.3 for more information about this procedure). As can be seen in the last row of table 2, these differences are not significant. The relations between the different forms and respectively age, sex and education are therefore in the same range.

As a third index of equivalence, the relations between the two versions and an external construct, namely the delayed recall score of the Rey-Osterrieth Complex Figure Test, are evaluated. When the two forms are representative to each other, the relation with this construct should not be significantly different. The relation between the two forms and the Rey-Osterrieth Complex Figure Test are not significant (r_s = 0.34, p = 0.14 for the original test; r_s = 0.36, p = 0.11 for the short form). Because these two correlations do not significantly differ from each other (z = 0.56, p = 0.58), it is concluded that the two versions are representative to each other with regard to their relation with the delayed recall score of the Rey-Osterrieth Complex Figure Test, an external construct.

At last, it is evaluated if the total scores of the two test versions can both equally predict the free recall scores. Since the Spearman's correlation coefficients (r_s = 0.87, p > 0.01 for the original test and the free recall score; r_s = 0.84, p > 0.01 for the short form and the free recall score) do not differ significantly (z = 1.44, p = 0.15), it is concluded that the predictive value of both test versions fall in the same range with regard to predicting the delayed recall score.

Table 2. *Spearman's correlation coefficients between the total scores of the two forms and respectively age, sex, and education and the difference between these correlations.*

	AVLT Total score trial 1 to 5	AVLT Total score trial 1 to 3	Difference between the correlations
Age	r_s = -0.57* (*p* = 0.01)	r_s = -0.60* (*p* = 0.01)	z = 0.74 (*p* = 0.46)
Sex	r_s = 0.38 (*p* = 0.09)	r_s = 0.34 (*p* = 0.14)	z = 0.95 (*p* = 0.35)
Education	r_s = 0.10 (*p* = 0.67)	r_s = 0.13 (*p* = 0.56)	z = 0.69 (*p* = 0.49)

** p < 0.05*

Consistency between scores

The consistency between the ranking positions of the scores on the two test versions are compared to each other using a Wilcoxon Signed Rank Test. The ranking variables of the total scores in both test versions do not differ significantly (*p* = 0.81), meaning that the ranks in the sample are assigned equally when using either the original test or the short form.

4.2.4 Conclusion and Discussion

General limitations of and recommendations based on this thesis can be found in paragraph 5.

The RAVLT can be shortened by presenting only the trials 1 to 3 instead of the trials 1 to 5. In the current study, the this three-trial short form has been compared with the original five-trial form of the RAVLT (Dutch version). The short form appears to be highly equivalent to the original form when their relation and the relations with demographic variables, an external construct, and the delayed recall score are taken into account. Further, there is consistency between the ranking scores of the two test versions. These findings suggest that the three-trial instead of the five-trial version can be used to obtain a direct recall score, which provides considerable time savings.

Although the three-trial version of the RAVLT is very promising, future research is needed. First, the influence of the shortening of the test on the scores for delayed recall and recognition are not determined in this study, since all participants were exposed to all five trials during the test administration. To assess the influence of a reduced amount of trials on the scores of delayed recall and recognition, a study is needed in which a large sample is divided into a group that takes the five-trial version and a second group, similar regarding demographic variables, that takes the three-trial version. In such a study, these influences can be evaluated. Second, the influence of the shortening of the RAVLT on the amount of repetitions and intrusions is also still unknown and could not be investigated in this study,

since the participants responded with almost no repetitions and intrusions. Therefore, this influence should be assessed in patient groups that are known to give repetitions and intrusions.

For the collection of normative data for the short form, the importance of monitoring the influence of demographic variables needs to be stressed. In this study, the relations between the immediate recall scores and respectively age, sex and level of education are assessed. This revealed only an influence of age. However, in other studies, more demographic variables have been found to affect test performance, such as age, ethnicity and education (Uchiyama et al., 1995). It is known that difference between the influence of demographic variables are found across several studies (see for a review van der Elst and colleagues, 2005).

Additionally, reported is that the influence of age on the RAVLT seems to be overestimated when the total scores are used to define performance (van der Elst et al., 2005). Suggested is the use of a learning measure that corrects for the performance on the first immediate recall trial, so that a more realistic estimation of the verbal learning capacity of an individual can be given. Moreover, van der Elst and colleagues (2005) suggest the use of a learning measure that takes ceiling effects into account, as this also contributes to a more realistic estimation. In this context, the use of a total score derived from trials 1 to 3 is more convenient than the use of a total score derived from trials 1 to 5, since theory suggests that the total score of the shorter version takes ceiling effects better into account (van der Elst et al., 2005). This provides another argument to use the three-trial version that is evaluated in this thesis instead of the original five-trial version.

The objective of the current study was to find a way to minimize the assessment time of the RAVLT. The three-trial version seems to be a good answer to this need. There is however yet another way to accomplish a reduced administration time. That is, when not the test itself but the procedure of administration is alternated. Zhao, Lv, Zhou, Hong and Guo (2012) investigated the influence of reducing the time between the first five subsequent trials (free recall) and the trials for delayed recall and recognition when detecting amnestic mild cognitive impairment. When comparing a 20 minute delay time with a 3 to 5 minute delay time, the difference in scores were so little that Zhao and colleagues (2012) concluded that a short delay time could substitute a longer delay time, especially for the oldest patients in their sample. No conclusion about other patient groups could be derived from this study, but it could be beneficial to use this shortened delay time, especially when the RAVLT is not part of a lager test battery and the delay time does actually delay the time in which a patient is occupied.

In conclusion, when the reduction of assessment time of the RAVLT itself is necessary, the three-trial version instead of the five-trial version can be used to measure the

immediate memory span. In addition, the three-trial version could even be seen as preferable above the five-trial version, since it takes ceiling effects better into account. It is therefore highly recommended to study the influence of the use of the three-trial version instead of the five-trial version on the delayed recall, the recognition, the repetitions and the intrusions. Although the investigation of all difference between the original form and the three-trial version and the collection of normative data for this shortened version demand much time and effort, it can certainly yield to profit. The RAVLT is, after all, very frequently used in NPE and the use of a three-trial version is very promising as a stand-alone measure.

4.3 Boston Naming Test

4.3.1 Literature Review and Introduction

The literature search revealed eleven articles about abbreviated forms of the Boston Naming Test (BNT) that were selected for further examination based on their titles and abstracts.

The BNT (Kaplan, Goodglass & Weintraub, 1983) originally consised of 85 items, but the current standard test consists of 60 black-and-white drawings that are ranked based on their familiarity. Participants have to name the items, starting with the 30th item. When they make a mistake at the first eight items, the test is continued in reversed order until eight consecutively correct responses are given. When patients are unable to name a drawing, a semantic cue is provided, which is followed by an phonological cue if necessary (Lezak et al., 2012). The test measures naming impairments, which can be caused by several pathologies such as aphasia, right hemisphere damage and Alzheimer's disease. A decline in naming ability is also a phenomenon that occurs in normal aging (Au et al., 1995). The BNT is the thirteenth most used neuropsychological test in the United States and Canada, according to a survey (Rabin, Barr & Burton, 2005).

The time that it takes to administer, score, interpret and report the BNT is approximately 23 minutes (Lundin & DeFilippis, 1999), and varies a lot between unimpaired and impaired patients. It is therefore not surprising that numerous short forms of the BNT have been developed. In the literature search, seven 30-items versions and six 15-items versions were revealed, together with two short forms based on item response theory (Graves, Bezeau, Fogarty & Blaire, 2004) and an empirically derived short form (Lansing, Ivnik, Cullum, & Randolph, 1999). Further, a 15-items version is developed for the testbattery of the well-known Consortium to Establish a Registry for Alzheimer's Diesase (CERAD; Morris et al., 1989). For two 30-items versions is concluded that they were essentially equivalent to the original version (Saxton et al., 2010).

For the Dutch version of the BNT (van Loon - Vervoorn, Stumpel, de Vries, 1995), a short form is also developed. This short version consists of 29 images that are also part of the original BNT (van Loon - Vervoorn, 2005). It is reviewed as equivalent to the full version of the BNT and the scores of the short form can be converted so that they can be compared with the normative data of the original 60-items test. For these reasons, this short form is frequently used in many clinical settings throughout the Netherlands. Furthermore, a 15-items version is recently developed for the use in clinical settings as well. Therefore, the aim of this study is to evaluate whether the quality of this 15-items version is equivalent to the quality of the 29-items version.

4.3.2 Methods

For the BNT, it will be evaluated whether the 15-items version is equivalent to the 29-items version. The Dutch BNT is assessed and all participants, of which more information can be found in paragraph 2.3, have Dutch as their first language. The 15-items version consists of items 1, 2, 5, 6, 9, 10, 13, 15, 17, 23, 24, 25, 27, 28 and 29 of the 29-items version. More information about these items and the items that are present in the 29-items version can be found in appendix 2.

The reliability, equivalence, consistency between scores for both versions and diagnostic agreement will be analyzed as described in paragraph 2.3. For the comparison between the relations of the two test versions with an external construct, the relations with the Token Test (21-items version) are used. When defining the diagnostic categories, the scores of the 29-items version will be converted using the formula (60/29)*score and the scores of the 15-items version will be converted using the formula (60/15)*score, so that the scores can be compared to the normative data that is available for the 60-items version (van Loon - Vervoorn, 2005; the control group that is used in this study is “other participants”).

4.3.3 Results

Reliability

The degree of consistency across items was assessed for both the 29-items form and the 15-items form of the BNT. This degree was adequate in the 29-item form, Cronbach’s $\alpha = 0.74$, and low in the 15-item form, Cronbach’s $\alpha = 0.57$.

Equivalence

To determine if the total score derived from the 15-items form is representative to the total score derived from the 29-items form, the relation between these two total scores are examined, whereby a relation close to 1 is considered as representative. Spearman’s correlation coefficient is used to define this relation, because the data are not normally distributed. The correlation is strong ($r_s = 0.78$, $p < 0.01$).

As a second index of equivalence, the relations between the original version and demographic variables (namely age, sex and education) are compared to the relations between the abbreviated form and these same variables. It is expected that the abbreviated form relate to these variables in the same way (parallelism). Firstly, the Spearman’s correlation coefficients are calculated. As can be seen in table 3, the level of education is significantly correlated to the scores of the BNT: Participants with an higher level of education are scoring higher on the two test forms than participants with a lower level of education ($r_s = 0.74$, $p < 0.01$ for 29-items form; $r_s = 0.55$, $p < 0.05$ for the 15-items form).

Table 3. *Spearman's correlation coefficients between the total scores of the two forms and respectively age, sex, and education and the difference between these correlations.*

	BNT 29-items form	BNT 15-items form	Difference between the correlations
Age	r_s = -0.20	r_s = 0.13	z = 2.13* (*p* = 0.03)
Sex	r_s = -0.43	r_s = -0.28	z = 1.03 (*p* = 0.30)
Education	r_s = 0.74**	r_s = 0.55*	z = 1.71 (*p* = 0.09)

** p < 0.05. ** p < 0.01.*

The correlations between the total scores and respectively age and sex are not significant in both forms of the BNT (see table 3). Secondly, the correlations of the two test versions with the demographic variables are compared to each other. The correlations with sex and level of education are in the same range (see table 3), but the correlation with age is different for the 29-items form and the 15-items form (z = 2.13, *p* < 0.05).

The correlations with an external construct are also compared to each other, namely the correlations with the scores on the Token Test (21-items version). The Spearman's correlation coefficient is not significant in both test forms (r_s = 0.20, *p* = 0.39 for the 29-items form; r_s = -0.02, *p* = 0.93 for the short form). Because the two correlations do not significantly differ from each other (z = 0.40, *p* = 0.16), it is concluded that the two versions are representative to each other with regard to their relation with the Token Test, an external construct.

Consistency between scores

The consistency between the ranking positions of the scores in the two test versions are compared to each other using a related-samples Wilcoxon Signed Rank Test. The ranking variables for the total scores on both test versions do not differ significantly (*p* = 0.93), meaning that the ranks in the sample are assigned equally in both the 29-items form and the 15-items form.

Diagnostic agreement

Due to ceiling effects, performance on the BNT can never be rated as 'excellent'. In the sample, all other categories were present for the scores on both test versions. The rate of agreement between the 29-items form of the BNT and the 15-items form of the BNT is moderate, κ = 0.53.

4.3.4 Conclusion and Discussion

General limitations of and recommendations based on this thesis can be found in paragraph 5.

In this study, it is evaluated whether the 15-items version of the Dutch BNT is equivalent to the 29-items version. The results show that the reliability of this 15-items version is, in contrast with the reliability of the 29-items version, low. With this degree of reliability, the 15-items version does not meet the criteria for psychological tests of the Dutch psychological association (COTAN; Egberink, Janssen & Vermeulen, 2014). In addition, a low reliability means that the test is not consistent and therefore not informative about the construct that should be measured (Goodwin, 2008). Furthermore, not all correlations of the two test forms with demographic variables where in the same range, indicating that the 15-items version is not representative to the 29-items version. It is therefore not surprising that the diagnostic agreement between the two test versions of the BNT is only moderate, which is not sufficient when the 15-items version is used as replacement of the 29-items version. For these three reasons, the evaluated 15-items version should not be used either as a replacement of the 29-items version in NPE nor as a screening instrument.

Although the evaluated 15-items version of the Dutch BNT emerges as inadequate, it is possible that other 15-items versions are adequate when measuring naming abilities. Future research could focus on finding an optimal short form of the BNT, since the BNT is used very often wherefore the reduction of its assessment time is valuable. When developing a short form in the Netherlands, differences in culture and language should be taken into account. Namely, the responses given to the items of the BNT show cultural difference (e.g. Fillenbaum, Huber & Taussig, 2008) and a Spanish version, the Texas Naming Test, has a greater sensitivity for Spanish Speakers than a translated version of the BNT (Marquez de la Plata et al., 2008 in Lezak et al., 2012). That is, a short form of the BNT that is reviewed as good in other language areas does not necessarily have to be reviewed as good in the Netherlands.

In sum, the 15-items version of the BNT that is reviewed in this study is not representative to the 29-items version and should not be used in NPE or for screening purposes. Notwithstanding, future research could find other short versions of the BNT that are superior to the 15-items version that is evaluated in the current study, whereby differences in quality of the BNT in different language and cultural areas should not be ignored.

4.4 Verbal Fluency, Phonological (Letter N and Letter A)

4.4.1 Literature Review and Introduction

During the literature search, seven articles were selected for further examination based on their titles and abstracts.

A popular test in neuropsychological assessment is the test of phonological verbal fluency or the Controlled Word Association Test, in which a patient has to produce as many words that begin with a certain letter as he or she can, within a small period of time. All words listed in the dictionary are correct, except proper names (such as names of people or places) and the same words with different endings (such as 'row' and 'rowing'). The test measures the speed and ease of verbal production, but also the ability to organize output (executive functioning) and sustained attention (Zakzanis, McDonald & Troyer, 2013). The Thurstone Word Fluency Test (1939), in which participants had to write down as many exemplars of one specific category as possible, took nine minutes and inspired Benton to create a three-minute version of this test, which eventually became part of a brief aphasia test battery, and received the name Controlled Word Assocation Test. This name was chosen to minimize confusion with the fluent or nonfluent dimension of aphasic speech (Benton & Hamsher, 1989; Ruff, Light & Parker, 1996), but the term verbal fluency is still commonly used when referring to this test.

Two methods can be applied to shorten the verbal fluency test. One method is to shorten the verbal fluency test by reducing the response time that is given for every trial (every letter). Usually, a response time of one minute is given, although participants sometimes get one and a half or even two minutes to respond. In the context of efficient neuropsychological assessment, an one minute response time is of course the most beneficial. The diagnostic utility of one minute fluency measures is evaluated for several pathologies, for example for Alzheimer disease and vascular dementia (Canning, Leach, Stuss, Ngo & Black, 2004). When semantic (category animals) and phonemic (only the letter F) fluency measures were combined, the one-minute test versions were able to discriminate between the two etiologies. Therefore, a response time of one minute seems to be not only beneficial for the assessment time, but also good enough for diagnostic purposes. In this thesis there will not be investigated whether the most frequently used response time of one minute can be reduced even further, since this will save only an insignificant amount of time.

A second method to shorten the verbal fluency test is to minimize the number of trials (i.e. first letters). In Great Britain, the differences in performance of a three-minute version (letters 'F', 'A', 'S') and an one-minute short version (only the letter 'B') was examined (Harrison, Buxton, Husain & Wise, 2000). The scores of the two versions correlated highly

and the test retest reliability was only somewhat higher for the three-letter version. It was therefore concluded that little advantage is obtained from administering a three-letter version of the verbal fluency test instead of an one letter-version. Further, participants showed both improvement as well as deterioration at retest, so practice does not necessarily lead to improvement. In the Netherlands, not a three-letter version but a shorter two-letter version with the starting letters 'N' and 'A' is often used. As far as known, no studies are yet published about an one-letter version of the phonological fluency test for the Dutch language. Since the English one-letter version seems to be a substantial replacement for the three-letter-version, the current study will examine whether the one-letter version is also a substantial replacement for the Dutch version of the phonological fluency test with two letters.

4.4.2 Methods

It will be examined if the Dutch verbal fluency test in which two first letters ('N' and 'A') are given is equivalent to a short form in which only one letter ('N' or 'A') is given. For this purpose, the scores derived from the two trials separately (one with starting letter 'N' and one with starting letter 'A') will be compared to the scores derived from two trials combined (starting letters 'N' and 'A'). So, two short forms are evaluated. To examine if practice effects influence the second trial (letter 'A') and therefore the total score, the scores on the two separate trials will be compared. When improvement on the second trial is evident, practice effects will be taken into account.

Reliability analyzes cannot be conducted, since items cannot be defined. The equivalence, consistency between scores and diagnostic agreement will be investigated as discussed in the general methods of this thesis (see paragraph 2.3). In addition, it will be evaluated which of the two short forms predicts the score of the total form the best by comparing the correlations between the two short forms and the total form. To evaluate whether the relations of the test versions with an external construct differ, the relations with the semantic category of the verbal fluency test (scores for two minutes) will be examined. When defining the diagnostic categories, normative data are used which points percentile scores to both the total test as to the scores on the two letters separately (Nys, van Zandvoort, de Haan, n.d.; Brands, Kessels, de Haan; n.d.).

4.4.3 Results

There is no improvement or decline in the scores on the second trial compared to the first trial: The mean for the first trial, letter 'N', is 14.05 (SD = 5.08) and the mean for the second trial, letter 'A', is 13.00 (SD = 4.36). In addition, 12 of the 21 participants had a lower score on

the second trial. A Wilcoxon Signed Rank Test shows that the difference between the two separate letters is not significant (p = 0.22). Therefore, practice effects or a decline in performance will not be taken into account.

Equivalence

Since the data are not normally distributed, Spearman's correlation coefficient is used to define the relations between the scores for the letter 'N', the scores for the letter 'A', and the total scores. There is a strong relation between the letter 'N' and the total score (r_s = 0.92, p < 0.01) and between the letter 'A' and the total score (r_s = 0.94, p < 0.01). This suggests that the scores on the two single letters are representative to the combined scores. Because for this test two short forms are evaluated, it is interesting to see whether both short forms can both equally predict the total score. Therefore, the relations are compared to each other. Since the Spearman's correlation coefficients do not differ significantly (z = 0.78, p = 0.44), it is concluded that both short forms predict the total score equally.

As a second index of equivalence, the relations between the original version and demographic variables (namely age, sex and education) are compared to the relations between the short forms and these same variables. It is expected that the scores relate to these variables in the same way (parallelism). First, Spearman's correlation coefficients are calculated. These show that only the level of education is significantly correlated to the verbal fluency measures, whereby the participants with an higher level of education perform better than participants with a lower level of education (r_s = 0.48, p < 0.05 for only the letter 'N', r_s = 0.50, p < 0.05 for only the letter 'A' and r_s = 0.52, p < 0.05 for the original form with both letters). The correlations between the total scores and respectively age and sex are not significant in all test versions (see table 4). As can be seen in the last two rows of table 4, there is no significant difference between the relations of the original form and the relations of the score for only letter 'N' with the three different variables. There is no significant difference between the relations of the original form and the score for only letter 'A' either. In other words, the relations of the short forms are in the same range as the relations of the original form.

As a third index of equivalence, the relations between the different versions and an external construct are considered, namely the relations with the semantic verbal fluency test (with a response time of two minutes). When the short forms are representative to the original form, the relations with this construct should not be significantly different. The Spearman's correlation coefficient is significant in all three versions (r_s = 0.67, p < 0.01 for the original test; r_s = 0.62, p < 0.01 for only the letter 'N' and r_s = 0.68, p < 0.01 for only the letter 'A'). These correlations do not differ significantly when the original form and only the letter N are compared (z = 0.64, p = 0.52). When the original form and only the letter 'A' are

Table 4. *Spearman's correlation coefficients between the total scores of the three forms and respectively age, sex, and education and the difference between these correlations.*

	Age	**Sex**	**Education**
Verbal Fluency letter N	r_s = 0.03	r_s = -0.032	r_s = 0.48*
Verbal Fluency letter A	r_s = -0.15	r_s = 0.01	r_s = 0.50*
Verbal Fluency letter N+A	r_s = -0.09	r_s = -0.01	r_s = 0.52*
Difference in correlations between version N & N+A	z = 1.26 (*p* = 0.21)	z = 0.25 (*p* = 0.80)	z = 0.41 (*p* = 0.68)
Difference in correlations between version A & N+A	z = 0.74 (*p* = 0.46)	z = 0.19 (*p* = 0.85)	z = 0.21 (*p* = 0.42)

** p < 0.05.*

compared, this correlation does not differ significant either (z = -0.22, *p* = 0.83). Therefore it is concluded that both short forms are representative to the original form with regard to their relation with the semantic category of the verbal fluency test, an external construct.

Consistency between scores

The consistency between the ranking positions of the scores in the different test versions are compared to each other using a related-samples Wilcoxon Signed Rank Test. The ranking variables of the total scores do not differ significantly when the letter 'N' and the total version are compared (*p* = 0.74). So, the ranks in the sample are assigned equally in both the original test as in the short form. The same conclusion applies for the consistency between the total version and the letter 'A', since the ranking variables of these test versions do not differ significantly either (*p* = 0.95).

Diagnostic Agreement

Using the normative data described in the method section of this test, diagnostic categories are assigned to the data of both the total test as the data of the two short forms (i.e. the two letters separately). The diagnostic categories 'excellent', 'above average', 'average' and 'below average' are assigned in all three test versions. The diagnostic category 'impaired' is never assigned. The rate of agreement between the total version of the phonological fluency and the short form with only the letter 'N' is moderate, κ = 0.51, as is the rate of agreement between the total version and the short form with only the letter 'A', κ = 0.56.

4.4.4 Conclusion and Discussion

General limitations of and recommendations based on this thesis can be found in paragraph 5.

The aim of this study was to investigate whether an one-letter version of the phonological verbal fluency test (either letter 'N' or 'A') can replace a two-letter version of this test. A previous study in Britain revealed a high relation between an one-letter version and a three-letter version (Harrison et al., 2000). In the current study, the relations between the short forms and the two-letter version are strong as well, which suggests that the short forms are representative to the two-letter version. Also, the relations of the short forms with demographic variables and an external construct are in the same range as the relations of the two-letter version with these variables, suggesting parallelism. Further, there is consistency between the ranking positions obtained in the different test versions. Although these findings suggest that the short forms are representative to the original form, the diagnostic agreement of both short forms with the original form is only moderate. For this reason, an inaccurate conclusion about the performance of a participant on the verbal fluency test will be reached too often when a short form instead of the two-letter version is used. The use of only one letter instead of two letters in the phonological fluency test is therefore not recommended.

However, for the one letter version in the English language, diagnostic utility is observed (e.g. Canning et al., 2004). It is possible that although the use of only one letter is not diagnostically equivalent to the use of two letters, the use of one letter still has diagnostic value. Still, there are three reasons why the use of this one-letter version instead of the two-letter version is not recommended. Firstly, although the one-letter versions could have diagnostic value, the current study shows that these values are not equivalent to the diagnostic value of the two-letter version. Secondly, it is known that test retest reliability is lower when only one letter is used (Harrison et al., 2000). Lastly, the time gain obtained from using an one-letter version instead of a two-letter version is almost insignificant. Although this reduction in assessment time can be an advantage, for example when patients become frustrated when they need to name words with yet another letter after their first attempt failed, reduction in assessment time also has disadvantages. Several examples are discussed in paragraph 5. For the phonological fluency test, it is important to keep in mind that, although the assessment time will be reduced by 50% if an one-letter version instead of the two-letter version is administered, it is still a time gain of only one minute. In conclusion, since the time gain is so small and the diagnostic quality of the one-letter versions has shown to be less than the diagnostic quality of the two-letter versions, it is recommended to use the two-letter version in neuropsychological assessment.

4.5 Verbal Fluency, Semantic (Animals)

4.5.1 Literature Review and Introduction

In the semantic verbal fluency test, patients have to name as many words as possible, just as in the phonological verbal fluency test. Where the words have to start with the same letter in the phonological fluency tests, the responses in the semantic verbal fluency tests do not have to start with the same letter but have to be in the same semantic category. The categories most often presented are animals and professions. Next to the assessment of speed and ease of verbal production, the fluency tests also asses the ability to organize output (executive functioning) and sustained attention (Zakzanis et al., 2013). Although the semantic verbal fluency test is sometimes seen as the same measure as the phonological verbal fluency test, there is a difference in brain regions that are involved in the performance on the two verbal fluency tests: the phonemic fluency depend on the activity in the frontal lobe region, whereas the semantic fluency depends not only on the frontal brain function, but more on the temporal brain function (Zakzanis et al., 2013). Therefore, in this thesis the semantic verbal fluency test is reviewed separately from the phonological fluency test.

The assessment time of the semantic fluency consists of a short instruction, together with a response time of one, one and a half or two minutes for the presented category. Sometimes, more than one category is presented (for example both animals and professions). In the context of efficient neuropsychological assessment, it is of course more beneficial to give a response time of only one minute and to present only one category. Harrison and colleagues (2000) compared the semantic fluency scores for a response time of one versus one and a half minute, and report that no particular advantage obtained from running the longer version. The test-retest reliability was however a little greater for the longer version. The diagnostic utility of the one-minute semantic fluency test, with only one category (animals), was investigated in a study with patients with beginning Alzheimer disease and beginning vasculair dementia: When this semantic fluency test was combined with a phonemic fluency measure, it was possible to discriminate between the two etiologies (Canning et al., 2004).

In clinical settings, often only the category animals is presented in neuropsychological assessment. However, the response time for this test varies between one and two minutes. Based on the presented literature, it is expected that the use of a response time of one minute results in only a small decline in psychometric quality compared to the two-minutes response time. It will be investigated in this study if the use of a response time of one minute is really equivalent to the use of a response time of two minutes.

4.5.2 Methods

In this study, the scores of the semantic fluency test (category animals) with a response time of one minute will be compared to the scores of the test with a response time of two minutes, to see if these test versions are representative to each other. First, it will be explored whether the performance in the first or in only the second minute differs, so that possible practice effects or a decline in functioning can be taken into account.

Reliability analyzes cannot be conducted, since items cannot be defined. The equivalence and consistency between scores will be investigated as discussed in the general methods of this thesis (see paragraph 2.3). To compare the difference in the relations with an external construct, the correlations between the two different scores and the score for the phonological verbal fluency (original test with both the letters N and A) will be compared.

The diagnostic agreement cannot be defined for the two response times of this test for two reasons. First, there is a significant difference between the amount of animals named in only the first and in only the second minute (related samples Wilcoxen Signed Rank Test, $p < 0.01$), so the scores obtained with a one-minute response time cannot be compared to the normative data of scores with a two-minutes response time when they are multiplied by two, since there is no linear relationship. The second reason is that no normative data are available to the researcher in which scores after both one and two minutes are considered. Using two different norm groups when defining the diagnostic agreement results possibly in a difference in diagnostic categories that is not caused by the difference in response time, but by the use of two different norm groups. Therefore, no conclusion will be made about the diagnostic agreement between the two response times.

4.5.3 Results

Data from one participant is excluded from the analyzes, since the one-minute score was not reported for this participant.

It is evaluated whether there is an improvement or decline in the second minute of the response time compared to the first minute. The mean number of named animals after one minute was 26.20 (SD = 6.83) and the mean number of animals named in only the second minute was 16.75 (SD = 7.83). The named animals in the second minute are significantly less (related samples Wilcoxen Signed Rank Test, $p < 0.01$). Only one participant named more animals in the second than in the first minute. This decline in functioning will be taken into account (for example when the diagnostic agreement is concerned, see for more information paragraph 4.5.2).

Equivalence

The relation between the scores after a response time of one minute and the scores after a response time of two minutes examined. Spearman's correlation coefficient is used, since the data for all scores are non-normally distributed. The relation between the one-minute version and the two-minute version is strong ($r_s = 0.89$, $p < 0.01$).

The relations between the original version and demographic variables (namely age, sex, and education) are compared to the relations between the short form and these same variables. It is expected that the short form relates to these variables in the same way (parallelism). First, the Spearman's correlation coefficients are calculated. As can be seen in table 5, only the relation between the semantic verbal fluency test and level of education is significant: Participants with an higher education level are scoring higher on the test than participants with a lower education level ($r_s = 0.75$, $p < 0.01$ for the original test, $r_s = 0.61$, $p < 0.01$ for the test version with the shortened response time). Second, the correlations of the short form (one minute) and the original test (two minutes) are evaluated. The correlations with age and level of education are in the same range for both test versions (see table 5), but the correlation between the original test and sex is different for the original test and the test version with only one minute response time ($z = 2.27$, $p = 0.02$).

The relations with an external construct are also compared to each other, specifically the correlations with the phonological verbal fluency test (full version with both the letters 'N' and 'A'). The Spearman's correlation coefficients are significant ($r_s = 0.60$, $p < 0.01$ for the original test; $r_s = 0.67$, $p = 0.01$ for the short form). Because the two correlations do not differ significantly from each other ($z = 0.83$, $p = 0.40$), it is concluded that the two versions are representative to each other with regard to their relation with the phonological verbal fluency test, an external construct.

Table 5. *Spearman's correlation coefficients between the total scores of the two forms and respectively age, sex, and education and the difference between these correlations.*

	Semantic Fluency Animals 1 minute	**Semantic Fluency Animals 2 minutes**	**Difference between the correlations**
Age	$r_s = -0.36$	$r_s = -0.32$	$z = 0.35$ ($p = 0.72$)
Sex	$r_s = 0.20$	$r_s = -0.05$	$z = 2.27$* ($p = 0.02$)
Education	$r_s = 0.61$**	$r_s = 0.75$**	$z = 1.83$ ($p = 0.07$)

** $p < 0.05$. ** $p < 0.01$.*

Consistency between scores

The consistency between the ranking positions of the test scores are compared to each other using a related-samples Wilcoxon Signed Rank Test. The ranking variables of the total scores on both test versions do not differ significantly (p = 0.74), meaning that the ranks in the sample are assigned equally in both the original test and the short form.

4.5.4 Conclusion and Discussion

General limitations of and recommendations based on this thesis can be found in paragraph 5.

The goal of the study about the semantic verbal fluency test was to investigate whether a response time of one minute could replace the response time of two minutes in neuropsychological assessment. It is assumed that performance doesn't change when participants know they have either one or two minutes to respond. The results show that participants name more animals in the first than in the second minute. To see whether the one-minute response time can replace the two-minutes response time in neuropsychological assessment for diagnostic purposes, the equivalence between the two test versions is evaluated in several ways. Results show that the ranks in the sample are assigned equally in the two test versions. Although the two test versions are strongly related to each other and the relations of the two test versions with an external construct are in the same range, not all relations with demographic variables are in the same range. Although there is no significant relation between sex and either the one-minute version and the second-minutes version of the semantic verbal fluency test, these relations differ significantly from each other. Since the two test versions are not representative to each other with regard to their relation with sex, it cannot be guaranteed that the two test versions measure the same construct and can be used interchangeable in neuropsychological assessment.

Because Harrison and colleagues (2000) report no particular advantage when applying a response time of one and a half instead of one minute, it would be very interesting for future research to focus on the diagnostic agreement of the two test versions. This could be studied by obtaining normative data from a group in which the scores after both one minute and two minutes is recorded, so that the diagnostic outcome could be compared with the data of one single norm group. Another method is to find a non-linear formula by which the scores after one minute can be transformed so that they can be compared to normative data that is available for the two-minutes test version. When the diagnostic agreement is studied, more clarity about the difference between the one-minute and two-minutes version of the semantic fluency test can be provided.

However, the current study provides evidence that the two test versions are not interchangeable regarding their relation with a demographic variable. It is therefore advised to use the two-minutes test version until more clarity about the diagnostic agreement can be provided. Moreover, although the short version provides a time gain of 50%, the time gain is still very small: only one minute. This small advantage of the use of the short version should be compared to the disadvantages of the reduction in assessment time, such as a smaller test-retest reliability (Harrison et al., 2000), the uncertainty about the diagnostic agreement, and the non-equivalence between the one-minute and two-minutes test version that the current study revealed. It is therefore recommended to use the two-minute version of the semantic verbal fluency test in neuropsychological assessment used for diagnostic purposes.

4.6 Trail Making Test

Literature Review

The Trail Making Test (TMT) was developed at the end of the Second World War out of several earlier versions of the test and was incorporated in a standard army test battery (Brown, Casey, Fisch & Neuringer, 1958). Nowadays, the TMT is commonly administered in neuropsychological assessment. According to a large sample survey in the United States and Canada, the TMT is the most frequently used neuropsychological instrument after the Wechsler Adult Intelligence Scale and the Wechsler Memory Scale (Rabin et al., 2005). The TMT is used to measure visual search, attention, mental flexibility, and motor function (Spreen & Strauss, 1991, p. 322). Part A of the TMT requires the connection of 25 encircled numbers in the proper order. In part B, 25 encircled letters and numbers have to be connected in alternating order (Spreen & Strauss, 1991, p. 323). To interpret the performance, the time that a participant needs to complete the test is used. The observations can be very valuable in neuropsychological assessment as well. According to a survey among 107 neuropsychologists, it takes about 14 minutes to administer, score, interpret and report version A and B of the TMT (Lundin & DeFilippis, 1999). The time required for the test administration itself varies from five to ten minutes (Spreen & Strauss, 1991, p. 325).

Although the titles and abstracts of 317 studies found in literature search were evaluated, no studies about abbreviation of the TMT were identified. Several parallel versions of the TMT are developed, but these are not only similar to the original TMT in the psychometric qualities and formats, but also in the number of items (Atkinson, Ryan, Kryza & Charette, 2011). Although no short forms were found in the literature search, a screening version of the test does exist and is published in the Montreal Cognitive Assessment (MoCA; Nasreddine et al., 2005). In this screening instrument, a small version of only part B of the TMT is printed. This version is however not sufficient to use for diagnostic instead of screening purposes.

For this thesis, it is chosen not to develop a short form of the TMT for diagnostic purposes for the following reason. Since the length of time that is necessary to complete the test is tracked, the test can be stopped after a certain period of time. More specifically, when the duration of the performance already results in an impaired score, there is no use to continue the test administration, except when the score is used to compare other scores with (for example to compare the score of part A of the TMT with the score of part B). That is, when the quantitative interpretation of the test results are considered. If observation seems informative, it can of course be decided to continue the test performance, even after an

impaired score has been reached. Using this method, it is not necessary that the TMT takes more time than required. The administration time can be cut-off when the qualitative results are clear, but when more quantitative results seem to be valuable the test administration can be continued. This creates an optimal way of doing efficient neuropsychological assessment.

4.7 Rey-Osterrieth Complex Figure Test

Literature Review

The Rey-Osterrieth Complex Figure Test (RCF), which was developed by Rey in 1941 and standardized by Osterrieth in 1944, is a widely used neuropsychological test for the evaluation of perceptual organization and visual memory (Lezak et al., 2012). In the RCF, patients are asked to draw a copy of a complex figure. Several scoring systems have been published, but most commonly used is a scoring method which divides the figures into 18 scorable units. In addition to this scoring system, the strategy and perceptual organization of patients can be evaluated by comparing the degree to which the figure was drawn in a conceptual, fragmented, or confused manner to the performance of healthy people (Lezak et al., 2012). After the copy trial, usually a recall trial follows to evaluate the visual memory of a patient. This recall trial can be immediately after the copy trial, after a delay, or both. At last, a recognition trial can follow. Each trial of the RCF takes approximately 10 minutes to complete (Shin, Park, Park, Seo, & Kwon, 2006). Because this results in a relatively long assessment time, especially when more trials are presented, it would be beneficial to shorten the duration of the RCF.

However, the literature research did reveal only several parallel forms (e.g. the Taylor Figure; Strauss & Spreen, 1990) and no parallel short forms. Therefore, several options to reduce the assessment time of the RCF will be discussed in this paragraph. First, it is of course beneficial to minimize the amount of trials of the RCF. Although it differs for every patient which trials are needed to reach a conclusion about his or her cognitive functioning, the delayed recall trial is evaluated as a better measurement of visual memory than the immediate recall trial (Strauss et al., 2006). So when using the RCF, experts should be critical if the immediate recall trial is necessary for the neuropsychological evaluation. The same critical evaluation should occur when considering the recognition trial, which is of course only valuable when the delayed recall trial does not give enough information about the different aspects of memory functioning. Note that different normative scores should be used when the amount of trials of the RCF change, since learning effects do occur (Strauss et al., 2006).

A second option to reduce the assessment time of the RCF is to think about the scorable units of the figure as items of any other test and to reduce the total number of items, hence the total number of scorable units. Be that as it may, it stands to reason that all the parts of the figure are interdependent, so that deleting one part changes the whole figure and therefore the original test. Therefore, a single change could lead to the fact that other parts of the figure are more easily to detect or more difficult to remember. For this reason, the

development of such a short form of the RCF is complicated and goes beyond the scope of this thesis. Nonetheless, it is interesting for future research to focus on reducing the assessment time of the RCF in this way. A new figure with less units can be developed and subsequently compared to the original RCF in terms of difficulty, equivalence and diagnostic agreement.

A third option to reduce the assessment time of the RCF is to monitor the time it takes to complete the test and to stop the test after a certain amount of time. It is however important for the development of a short form that this short form measures the same construct as the original test. When adding a time variable to the RCF, it is likely that other constructs interact with the score, such as the speed of information processing. Therefore, this method to shorten the assessment time of the RCF does not seem advantageous and will not be further evaluated in this study.

At last, for the efficiency of neuropsychological assessment it is important to consider if the RCF should be really assessed. Namely, for some patients the RCF is too difficult to complete, for example for patients with severe visual-spatial problems or neglect. If this is a probability, it should be considered to present other easier figures, because these take less time to assess. Examples of such figures are a cube or other figures that are used in screening instruments (see for example Jonker & Lindeboom, 1989). After an easier figure is copied by a patient, it will appear whether it is useful to administer the RCF additionally or if enough information is already collected within a smaller length of time. This consideration should of course be made for each patient individually and relies on the expert opinion of the neuropsychologist.

In sum, the assessment of the RCF is very time consuming. Future research could focus on the possibilities to develop a figure that is representative to the original RCF but takes less time to copy. Currently, it is the task of the neuropsychologist to consider whether all trials of the RCF are, and even whether the RCF itself is really needed in the assessment.

4.8 Brixton Spatial Anticipation Test

4.8.1 Literature Review and Introduction

The Brixton Spatial Anticipation Test (BSAT), part of the Hayling and Brixton tests (Burgess & Shallice, 1997), is developed to assess executive functions, in particular rule detecting, impulsivity, and switching (Chan, Shum, Touloupoulou & Chen, 2008; Van Den Berg et al., 2009). The test consists of 56 pages. On each page, 10 circles are presented of which one circle is colored. It changes per page which circle is colored according to several patterns. Patients have to discover these patterns which change after several pages without warning and predict which circle is colored on the next page. The outcome measure is the total number of errors. The test can be compared to the Wisconsin Card Sorting Test, but the BSAT is better usable in the clinic because it is less-time consuming and less stressful than the Wisconsin Card Sorting Test (Chan et al., 2008). The administration time of the BSAT is approximately 10 minutes on average (Van Den Berg et al., 2009).

The literature search did not reveal any short form of the BSAT, possibly because the test is relatively new. Since the BSAT does not have a time restriction or cutoff point during administration, it could be very beneficial to use a short form. Therefore, in the current study a short form is developed. Subsequently, it's psychometric and diagnostic quality will be compared to the original BSAT.

4.8.2 Methods

The first part of this study is to develop a short form. It is chosen to reduce the original BSAT with approximately half of the items, so that the time gain that arises when using the short form is really significant. The BSAT consists of 55 items, but these items are divided into 9 series of different patterns. The first series can be viewed as the easiest series and are probably meant to be an introduction to the test or as practice series. Therefore, the first series will be left unchanged in the short form, so that the shortened version is as representative to the original test as possible. The other series will be ranged based on their difficulty. This ranking occurs by determining the total score per series in the sample and dividing these total scores by the number of items of which the series consist. After the ranking is determined, the eight series will be halved based on their ranking. That is, the most easy series are the first and as discussed these are probably meant for practice, so this series will be part of the short form. The second most easy series will not become part of the short form, but the third most easy series will. Ensuing, the fourth, sixth and eighth most easy series will not become part of the short form, but the fifth, seventh and ninth most easy series

will become part of the short form. More details about these series can be found in table 6. In this table, it can also be seen that the second and third rank are shared for the third and last series of the original BSAT. It is chosen to submit the last series to the short form and not the third, so that the short forms ends in the same way as the original test. The result is a short form with the series numbers 1, 3, 5, 8, 9. In addition, note that the number of items per series in the short form are representative to the number of items per series in the original form: Both small and large numbers of items per series are present in the short form.

To evaluate the short form, the reliability, the equivalence, the consistency between scores for both versions and the diagnostic agreement will be analyzed as described in paragraph 2.3. When comparing the relations with an external construct, the relations of the two versions with the Test of Everyday Attention, subtest Visual Elevator (timing scores of the original version) are compared to each other. When defining the diagnostic agreement, the scores on the short form are multiplied by a factor that is equal to the total number of items in the original test (55) divided by the total number of items in the short form (29). In this way, the scores on the short form can be compared to the normative data for the original BSAT. For every participant the total number of errors are transformed into percentile equivalents using table D of the scoring forms of the BSAT. These percentile equivalents are then converted to the diagnostic criteria defined by Lezak and colleagues (2012; see paragraph 2.3 of this thesis for the procedure).

Table 6. *Information about the series of the BSAT. Series included in the short form are in bold face.*

Original position of the series	**Items that are part of the series**	**Ranking position based on difficulty***
1	**2-6**	**1**
2	7-12	4
3	**13-19**	**5**
4	20-26	2/3
5	**27-29**	**9**
6	30-34	6
7	35-41	8
8	**42-48**	**7**
9	**49-55**	**2/3**

** The most easy series in the sample are indicated with the lowest number (1).*

4.8.3 Results

Data of one participant was not available for the BSAT.

Reliability

The degree of consistency across items was assessed for both the original BSAT as the short form. This degree was adequate in both the original test, Cronbach's α = 0.73, as in the short form, Cronbach's α = 0.71.

Equivalence

To determine if the total score derived from the original test is representative to the total score derived from the short form, the relation between these two total scores are examined. Spearman's correlation coefficient is used to define this relation, because the data are not normally distributed. The correlation is strong (r_s = 0.80, p < 0.01).

As a second index of equivalence, the relations between the original version and demographic variables (namely age, sex and education) are compared to the relations between the short form and these same variables. It is expected that the short form relates to these variables in the same way (parallelism). Although younger participants score better than older participants in only the original BSAT (r_s = -0.56, p < 0.05) and not in the short form (r_s = -0.38, p = 0.10), the correlations do not differ significantly (z = 1.40. p = 0.16). As can be seen in table 7, the relations between the two tests and respectively sex and level of education do not differ significantly either. This indicates that there is parallelism.

The correlations with an external construct are also compared to each other, namely the correlations with the Visual Elevator, a subtest of the Test of Everyday Attention (timing scores on the original test version). The Spearman's correlation coefficient is not significant for both test forms (r_s = -0.32, p = 0.17 for the original BSAT; r_s = -0.19, p = 0.42 for the short form). Because the two correlations do not significantly differ from each other (z = 0.89, p = 0.37), it is concluded that the two versions are representative to each other with regard to their relation with the score on the Visual Elevator, an external construct.

Consistency between scores

The consistency between the ranking positions for the scores on the two test versions are compared to each other using a related-samples Wilcoxon Signed Rank Test. The ranking variables for the total scores on both test versions do not differ significantly (p = 0.71), meaning that the ranks in the sample are assigned equally for both the original BSAT and the short form.

Table 7. *Spearman's correlation coefficients between the total scores of the two forms and respectively age, sex, and education and the difference between these correlations.*

	BSAT Original form	BSAT Short form	Difference between the correlations
Age	r_s = -0.56*	r_s = -0.38	z = 1.40 (*p* = 0.16)
Sex	r_s = 0.00	r_s = 0.12	z = 0.76 (*p* = 0.45)
Education	r_s = 0.26	r_s = 0.28	z = 0.12 (*p* = 0.91)

** p < 0.05.*

Diagnostic agreement

The performance of both test versions in the sample was rated as 'excellent', 'above average', and 'average'. The categories 'below average' and 'impaired' were not present in the sample. The rate of agreement between the original BSAT and the developed short form is fair, κ = 0.26.

4.8.4 Conclusion and Discussion

General limitations of and recommendations based on this thesis can be found in paragraph 5.

The goal of this study was to develop a short form of de BSAT that is representative to the original test, so that time can be saved during neuropsychological assessment. The short form is developed by arranging the series of items based on their level of difficulty that was observed in the sample. Then, half of the series of items was removed, so that both easy and difficult series of items are present in the short form. An exception is made for the first series of items, that remained present in the short form as well. In the developed short form are both small and large numbers of items per series present.

Because the short form was developed to replace the original BSAT in neuropsychological assessment that is used for diagnostic purposes, the short form should be representative to the original BSAT, which was evaluated in several ways. In terms of reliability, both the original and the short BSAT were adequate. Further, the versions had a strong relation to each other and the relations of the two variables with demographic variables and an external construct were in the same range, suggesting that the short form is equivalent to the original BSAT in this regard. In addition, the ranking positions in the sample are assigned equally for both the original BSAT and the short form, which suggests that the short form is representative to the original BSAT. However, examining the differences between the obtained scores in both test versions in more detail, suggests otherwise. Namely, the rate of agreement between both test versions is only fair, which is not enough

when the short form is used in neuropsychological assessment to obtain diagnostic conclusions. In conclusion, the developed short form of the BSAT is not equivalent to the original test when the diagnostic agreement is considered and the short form should therefore not be used for diagnostic purposes.

This is the first short version of the BSAT that is known to be developed. Because the level of difficulty of the series of items are taken into account when developing this short form, it is highly unlikely that another short form of the BSAT can be developed that has a better diagnostic agreement with the original BSAT. The focus of feature research should therefore not be on developing another short form of the BSAT, but on evaluating which test is most efficient when evaluating the executive functioning in neuropsychological assessment. With regard to the time gain, the use of the BSAT seems to be already a better choice than for example the use of the Wisconsin Card Sorting Test, which takes much longer to administer. Possibly, other efficient tests are available or could be developed.

In sum, the developed short form should not be used for diagnostic purposes. A suggestion for future research is to focus on the efficiency and duration of tests that evaluate the executive functioning, which could provide options for more efficient neuropsychological assessment.

4.9 Judgment of Line Orientation

4.9.1 Literature Review and Introduction

Ten relevant studies were obtained in the literature search.

In 1978, the Judgment of Line Orientation (JLO) was developed to measure visual-spatial perception (Benton, Hamsher, Varney & Spreen, 1983). The JLO is nowadays a very popular test in neuropsychological assessment (Rabin et al., 2005), as the test does not make an appeal to motor ability (Spencer et al., 2013). In the JLO, two short angled lines are presented, together with a set of (longer) lines with different angles. Patients have to decide for which lines the angle matches. The test knows 5 practice items and originally 30 items, and the scoring is based on the number of correct answers. According to a survey among 43 neuropsychologist, it takes approximately 21 minutes to administer, score, interpret and report the JLO (Lundin & DeFilippis, 1999). In addition, for severely impaired brain-damaged patients, it can be very difficult to complete the items, so the JLO can be even more time consuming to administer (Qualls, Bliwise & Stringer, 2010). It is therefore not surprising that several short forms of the JLO are developed.

Odd-items and even-item forms are developed (Woodard et al., 1996), since the original JLO has a good split-half reliability and the item difficulty of the JLO is designed to increase with each successive item (Benton et al., 1983). The validity of these short forms (the correlation with the original JLO) is defined as sufficient in a group of healthy older adults (Woodard, 1998), in a mixed clinical sample (Woodard et al., 1996; VanderPloeg, Lalone, Greblo & Schinka, 1997), and in a sample of patients with traumatic brain injury (Mount, Hogg & Johnstone, 2002). Sensitivity is defined as sufficient in a mixed clinical sample (Woodard et al., 1996). Even though these results suggest that the odd-item forms and even-item forms are a viable alternative to the full version of the JLO, not all studies support this point of view. Restrictions are a large margin of error (20% of all items; Woodard et al., 1996; Calamia, Markon, Denburg & Tranel, 2011), the distribution of correct responses (Qualls et al., 2010), and a reliability that does not met the recommended alpha value of 0.80 (Nunnally, 1978 in Winegarden, Yates, Moses, Benton & Faustman, 1998).

To develop an optimal short form of the JLO, Winegarden and colleagues (1998) evaluated the test results of a sample with neurologic, psychiatric and mixed diagnoses. Five short forms were evaluated, namely the odd-item and even-item forms, V1-10, V1-20 and V11-30. Based on the internal consistency and correlation with the full version, they recommend the use of the items V11-V30 as a short form.

In clinical settings, a short form with the items V10 to V24 is often used. Because this version consists of 15 items, it saves more time to use this version than to use the V11-V30

version with 20 items. The goal of this study is to evaluate whether the short form best reported in literature (V11 to V30) is representative to the original JLO and whether the short form often used in clinical settings (V10 to V24) is representative to the original JLO. In addition, it will be evaluated which short form is preferable for the use in clinical settings for diagnostic purposes.

4.9.2 Methods

The reliability, equivalence, the consistency between scores and the diagnostic agreement will be analyzed as described in paragraph 2.3. To compare the relations of the test versions and an external construct, the relations with the Rey-Osterrieth Complex Figure Test (copy trial) are considered. When defining the diagnostic agreement, the scores of the short forms are multiplied by 2 (version V10-24) or multiplied by 1,5 (version V11-30), so that the scores can be compared to the normative data that is available for the original JLO. The normative data that is used to assign diagnostic categories differs for every age group (Bouma, Mulder, Lindeboom & Schmand, 2012).

4.9.3 Results

Reliability

The reliability is examined by calculating the inter-item correlations. The consistency across items was adequate in both the original version V, Cronbach's α = 0.75, and in the form with items V11-30, Cronbach's α = 0.72. However, this consistency was marginal in the form with items V10-24, Cronbach's α = 0.61.

Equivalence

To determine if the short forms are representative to the original form, the relations between the short forms and the original version are examined, whereby relations close to 1 are considered as representative. Since the data are not normally distributed, Spearman's correlation coefficients are used to define these relations. The relation between V11-30 and the original version V was strong (r_s = 0.99, p < 0.01), just as the relation between V10-25 and the original version V of the JLO (r_s = 0.89, p < 0.01). Because for this test two short forms are evaluated, it is interesting to see whether both short forms are equally representative to the original test. That is, if they can both equally predict the total score of the original JLO. Therefore, the correlations of the two short forms with the original form are compared to each other. The Spearman's correlation coefficients do differ significantly (z =

5.51, $p < 0.01$). The short form with items V11-30 is more representative to the original test than the short form with items V10-25.

As a second index of equivalence, the relations between the original version and demographic variables (namely age, sex and education) are compared to the relations between the abbreviated forms and these same variables. It is expected that the abbreviated forms relate to these variables in the same way (parallelism). The correlations between these demographic variables and the three test versions can be seen in table 8. Sex is significantly correlated with the scores on the JLO, whereby man performed significantly better on all three forms ($r_s = -0.45$, $p < 0.05$ for the total form; $r_s = -0.45$, $p < 0.05$ for V11-30; $r_s = -0.48$, $p < 0.05$ for V10-24). The correlations between the levels of education and the three different forms are significant as well ($r_s = 0.48$, $p < 0.05$ for the total form; $r_s = 0.53$, $p < 0.05$ for V11-30; $r_s = 0.52$, $p < 0.05$ for V10-24). Of importance for this study is whether the correlations of the test versions differ. As can be seen in the last two rows of 9, there is no significant difference between the relations with the three demographic variables when the original JLO and the short version with V11-30 are compared. Neither is there a significant difference between these relations when the original JLO is compared to the V10-24 form. In other words, the relations of the short forms with demographic variables are in the same range as these relations of the original form.

As a third index of equivalence, the relations between the different versions and an external construct are considered, namely the score for the copy of the Rey-Osterrieth Complex Figure Test. When the short forms are representative to the original form, the relations with this construct should not be significantly different. The Spearman's correlation coefficient is significant in all three versions ($r_s = 0.41$, $p = 0.07$ for the original test; $r_s = 0.41$, $p = 0.07$ for the V11-30 version, $r_s = 0.30$, $p = 0.19$ for the V10-24 version). These correlations do not differ significantly when the original form and the V11-30 version are compared ($z = 0.00$, $p = 1.00$). When the original form and the V10-24 version are compared, this correlation does not differ significantly either ($z = 1.05$, $p = 0.29$). Therefore it is concluded that both short forms are representative to the original form with regard to their relation with the copy trial of the Rey-Osterrieth Complex Figure, an external construct.

Consistency between scores

The consistency between the ranking positions of the test scores are compared to each other using a related-samples Wilcoxon Signed Rank Test. The ranking variables of the total scores do not differ significantly when the original JLO and the V11-30 form are compared ($p = 0.95$), meaning that the ranks in the sample are assigned equally in both the original test and the V11-30 form. The same conclusion applies for the consistency between the total

Table 8. *Spearman's correlation coefficients between the total scores of the three forms and respectively age, sex, and education and the difference between these correlations.*

	Age	Sex	Education
JLO V TOTAL	r_s = -0.39	r_s = -0.45*	r_s = 0.48*
JLO V 11-30	r_s = -0.43	r_s = -0.45*	r_s = 0.53*
JLO V 10-24	r_s = -0.38	r_s = -0.48*	r_s = 0.52*
Difference in correlations between V 11-30 & V Total	z = 1.20 (*p* = 0.23)	z = 0.38 (*p* = 0.70)	z = 1.55 (*p* = 0.12)
Difference in correlations between V 10-24 & V Total	z = -0.07 (*p* = 0.95)	z = 0.19 (*p* = 0.85)	z = 0.42 (*p* = 0.67)

** p < 0.05.*

version and the V10-24 form, since the ranking variables of these test versions do not differ significantly either (*p* = 0.77).

Diagnostic Agreement

Due to ceiling effects, the classification 'excellent' can never be assigned to a score of the JLO. In the sample, the performance of all participants was not rated as impaired. Therefore, the diagnostic agreement was investigated by comparing the classifications 'below average', 'average', and 'above average'. The rate of agreement between the original version V of the JLO and the form with items V11-30 is almost perfect, κ = 0.818. The rate of agreement is only substantial for the original version V and items V10-24, κ = 0.724.

4.9.4 Conclusion and Discussion

General limitations of and recommendations based on this thesis can be found in paragraph 5.

In this study, two short forms are compared with the original JLO: The short form consisting of items V11-30 and the short form consisting of items V10-24. The literature review revealed that the V11-30 version is a promising fixed form for the use of a shortened stand-alone measure (Winegarden et al., 1998). In the current study, the inter-item reliability of this version shows to be sufficient, just as the inter-item reliability of the original test. This corresponded with the value of reliability found in earlier studies about this short form (Winegarden et al, 1998; Spencer et al., 2013). The current study also shows that the scores in this form are highly representative to the original test, and that the relationships of the original test and the V11-30 form with demographic variables and an external construct are in the same range. In addition, there is consistency between the ranks assigned to the scores

of both test versions. In previous studies, the diagnostic agreement of the V11-30 version was only defined as the difference between unimpaired and impaired scores (Spencer et al., 2013). The current study takes all diagnostic categories into account and appoints the rate of agreement as almost perfect. Summarily, the V11-30 version is representative to the original JLO and can be used a stand-alone measure.

The V10-24 version of the JLO, a short form often used in clinical settings, is also evaluated. Unfortunately, these results are less reassuring. The reliability is only marginal, and although the relation with the original JLO is strong, it is significantly less strong than the relation between the V11-30 version and the original JLO. There is equivalence between the V10-24 version and the original JLO with regard to their relations with demographic variables and an external construct and there is consistency between the ranks assigned to the scores of both test versions. However, the diagnostic agreement between the V10-24 version and original JLO is only substantial, which is lower that the diagnostic agreement between the V11-30 version and original JLO. In conclusion, the V10-24 version cannot be used as a replacement of the original JLO when diagnosing because of the degree of diagnostic agreement. Moreover, the marginal reliability makes that this 15-item version cannot be recommended for the use as a screening tool either.

The insufficiency of the reviewed V10-24 version but also the insufficiency of the odd-item and even-items forms can possibly be explained by the findings of two recent studies: Although Benton and colleagues designed the JLO so that the item difficulty increases with each successive item (1983), the items of the JLO do not really ascend in order of difficulty (Qualls et al., 2010; Calamia et al., 2011). Therefore, flexible short forms are developed that take the real item difficulty into account. The most promising is the flexible form developed by Calamia and colleagues (2011), in which item response theory (IRT) is used to determine the actual item's difficulty, and created a flexible form with basal and ceiling rules. When a patient correctly answers a predetermined quantity of items, it is assumed that he could answer the more easy items correctly as well. If a patient answers a predetermined quantity of items incorrectly, the test administration is stopped and it is assumed that the patient could not answer the more difficult items correctly either. This method of assessment results in an average of 20.4 items (SD = 5.4), and little difference between the score in the short form and original JLO. So, this flexible form seems very beneficial for the use as a stand-alone measure when diagnosing.

In conclusion, the current study reveals that the V11-30 version is representative to the original JLO and can be used as a stand-alone measure, whereas the use of the V10-24 version cannot even be recommended for the use as a screening tool. The V11-30 version was reported in earlier studies as the most promising fixed short form of the JLO. In addition, several flexible short forms are reported of which the form based on IRT is the most

promising (Calamia et al., 2011). For future research, it is recommended to compare these short forms, preferable by presenting the short forms beforehand, without reordering the items post hoc. In this way, possible influences of fatigue on the one hand or practice effects on the other hand can be taken into account when deciding about the most optimal short form of the JLO. Since in both promising short forms 20 items (on average) are assessed, another recommendation for future research is to find the most optimal 15-item short form of the JLO, since it saves time to assess 15 instead of 20 items. When developing this short form, it could be beneficial to focus on the level of difficulty of the items, since this level seems to be different than was thought originally (Qualls et al., 2010; Calamia et al., 2011) and a promising flexible form is already developed by focusing on this level of difficulty (Calamia et al., 2011). In any case, it is clear that there are short forms of the JLO that can be used to replace the original test, so that the relatively long assessment time of 21 minutes can be shortened and neuropsychological assessment can be conducted in a more efficient manner.

4.10 Token Test

4.10.1 Literature Review and Introduction

In 1962, De Renzi and Vignolo designed the Token Test to assess auditory comprehension in aphasics. In the original test, 61 verbal commands with increasing difficulty are given about 20 tokens in two different shapes, two different sizes and five different colors. In response, participants should provide the right gestures, such as pointing to or moving the tokens. The Token Test is popular among neuropsychologists: In 1991 it was used in neuropsychological assessment 61% of the time (Butler, Retzlaff & Vanderploeg, 1991). The original Token Test is very sensitive to disruptions in verbal comprehension (De Renzi & Vignolo, 1962), but it takes approximately 20 minutes to administer, score, interpret, and report the test (Lundin & DeFilippis, 1999). Since the publication of the Token Test, several short forms are presented (Spreen & Strauss, 1991, p. 268).

The literature search revealed 14 short versions of the English Token Test. The number of items of the short versions varied from 10 to 39. It goes beyond the scope of this thesis to review the quality of all short versions, for an overview is referenced to Lezak and collegues (2012). In general, the diagnostic agreement declines when less items are assessed, but this agreement is still sufficient in some short forms, for example when only the fifth and last part of the original Token Test is assessed. This fifth part consists of 22-items involving relational concepts and identified only one fewer patient as 'latent aphasic' than did the 62-items original test (Lezak et al., 2012).

In The Netherlands, a 21-items Token Test is often used (van Dongen, van Harskamp & Luteijn, 1976). This version consists of a combination of items from the fourth and fifth part of the original Token Test. It is seen as equivalent to the full version of the Token Test. To see whether this form can be shortened even more, this study will investigate whether half forms of this 21-items Token Test are equivalent to the 21-items Token Test, by evaluating the half even form and half oneven form of this test version.

4.10.2 Methods

For the Token Test, it will be evaluated whether the split-half versions are equivalent to the 21-items version. The Dutch version of the Token Test is assessed and all participants, of which more information can be found in paragraph 2.3, have Dutch as their first language. Since the 21-items version consists of an uneven number of items, the half versions differ in their amount of items. That is, the even version consists of 10 items, while the uneven

version consists of 11 items. More information about the items in the 21-items Token Test can be found in appendix 3.

The inter-item reliability, the equivalence, the consistency between scores for both versions and the diagnostic agreement will be analyzed as described in paragraph 2.3. To compare the relations with an external construct, the relations with the Boston Naming Test (29-items version) are used. When defining the diagnostic categories, the normative data for the 21-items form is used (van Dongen, van Harskamp & Luteijn, 1976). The scores of the half forms are converted so that they can be compared to the same normative data using the formula (21/10)*score for the even half form and (21/11)*score for the uneven half form.

4.10.3 Results

Reliability

The degree of consistency across items was assessed for all three test versions. This degree was adequate in the 21-items form, Cronbach's α = 0.72, marginal in the uneven form, Cronbach's α = 0.67, and low in the even form, Cronbach's α = 0.24.

Equivalence

To determine if the total scores derived from the half forms are representative to the total scores derived from the 21-items form, the relation between the different total scores are examined. Spearman's correlation coefficients are used to define these relations, because the data are not normally distributed. The correlation between the 21-items version and the uneven form is strong (r_s = 0.90, p < 0.01), just as the correlation between the 21-items version and the even form (r_s = 0.91, p < 0.01). This suggests that both half forms are representative to the 21-items form. Because for this test two short forms are evaluated, it is interesting to see whether both short forms can both equally predict the score of the 21-items form. Therefore, these correlations are compared to each other. Since the Spearman's correlation coefficients do not differ significantly (z = 0.24, p = 0.81), it is concluded that both test versions can equally predict the scores on the 21-items Token Test.

As a second index of equivalence, the relations between the 21-items form and demographic variables (namely age, sex and education) are compared to the relations between the abbreviated forms and these same demographic variables. It is expected that the abbreviated forms relate to these variables in the same way as the 21-items form (parallelism). First, the Spearman's correlation coefficients were calculated. The relations with sex and level of education are not significant (see table 9), age is only significantly related to 21-items and even form: Younger participants score better on these test forms

Table 9. *Spearman's correlation coefficients between the total scores of the two forms and respectively age, sex, and education and the difference between these correlations.*

	Age	**Sex**	**Education**
Token Test 21-item	r_s = -0.50*	r_s = 0.24	r_s = 0.19
Token Test Even	r_s = -0.57**	r_s = 0.15	r_s = 0.14
Token Test Uneven	r_s = -0.40	r_s = 0.21	r_s = 0.30
Difference in correlations: 21-items versus even form	z = 0.84 (*p* = 0.40)	z = 0.92 (*p* = 0.36)	z = 0.51 (*p* = 0.61)
Difference in correlations: 21-items versus uneven form	z = 1.07 (*p* = 0.28)	z = 0.29 (*p* = 0.77)	z = 1.08 (*p* = 0.28)

*** p < 0.01.*

than older participants (r_s = -0.50, $p < 0.05$ for the 21-items form; r_s = 0.57, $p < 0.01$ for the even form). As can be seen in the last two rows of table 9, the relations between the different test versions and respectively age, sex, and level of education do not differ significantly from each other. The relations of the two short form and the demographic variables are in the same range as the relations of the 21-items form and these demographic variables.

The correlations with an external construct are also compared to each other, namely the correlations with the score on the Boston Naming Test (29-items version). The Spearman's correlation coefficients are not significant for all three test forms (r_s = 0.20, p = 0.39 for the 21-items form; r_s = 0.09, p = 0.71 for the even half form; r_s = 0.35, p = 0.12 for the uneven half form). Moreover, the correlations do not differ significantly when the original form and the even half form are compared (z = 1.11, p = 0.27). When the original form and the uneven half form are compared, these correlations do not differ significant either (z = 0.99, p = 0.32). Therefore it is concluded that both short forms are representative to the original form with regard to their relation with the Boston Naming Test, an external construct.

Consistency between scores

The consistency between the ranking positions for the scores on the different test versions are compared to each other using a related-samples Wilcoxon Signed Rank Test. The ranking variables for the total scores do not differ significantly when the 21-items version and the even half form are compared (p = 0.79), meaning that the ranks in the sample are assigned equally for both the original test and the short form. The same conclusion applies for the consistency between the 21-items version and the uneven half form, since the ranking variables for these test versions do not differ significantly either (p = 0.14).

Diagnostic agreement

Due to ceiling effects, performance on the Token Test can never be rated as ‘excellent’. In the sample, all other categories were assigned to the scores, although the category ‘impaired’ was present only once in the uneven short form. The rate of agreement between the 21-items version and the even half form is moderate, $\kappa = 0.57$, and the rate of agreement between the 21-items version and the uneven short form is substantial , $\kappa = 0.76$.

4.10.4 Conclusion and Discussion

The goal of this study was to evaluate whether the 21-items version of the Token Test, a short version that is often used in clinical settings and reviewed as representative to the original 61-items form, can be shortened even more. For this research goal, the even half form and uneven half form are compared to the 21-items version of the Token Test. Note that the difference between the number of items of the two half forms (10 versus 11) is highly unlikely to cause any substantial differences between the quality of the half forms.

Both half forms have a strong relation with the 21-items version. Further, the relations of the 21-items version with demographic variables and an external construct are in the same range as these relations in both half forms. The ranking positions within the sample are also assigned equally when the 21-items form is compared to both half forms separately. However, the diagnostic agreement with the even form and the 21-items form is only moderate, so that it is concluded that this form is not sufficient for the use in neuropsychological assessment. Although the diagnostic agreement with the uneven form and the 21-items form is a little higher, namely substantial, this degree of agreement is still too low for the uneven form to replace the 21-items form when diagnosing. In addition, the reliability of both forms is not sufficient, so that the short forms cannot even be used as screening instruments.

Although these half forms are not convenient in neuropsychological assessment, it is possible that other short forms can be derived from the 21-items form, for example by focusing on the item difficulty in aphasics. It is however also possible that the 21-items form is the most efficient short form of the Token Test that can be used for diagnostic purposes, since this form is already an abbreviation of the original 61-items form. The omission of 40 items without losing much psychometric and diagnostic quality is remarkable. The current study emphasizes the difficulty of producing an even shorter form than the 21-items form without losing too much quality. Presumably, the limit of justified shortening of the Token Test is reached.

General limitations of and recommendations based on this thesis can be found in paragraph 5. However, one of this limitations needs to be stressed in the context of this specific test. In this thesis, the representativeness of the short forms with the original tests is only compared in a healthy control group. Especially for tests with ceiling effects such as the Token Test, it is important to test the diagnostic agreement and psychometric qualities of a short form within patient groups as well. Large differences in responses can occur when patients and not healthy adults are tested, which can result in large differences in the rating of the psychometric and diagnostic qualities. Such extensive research has been done for the 21-items test, which is, according to the literature review and the current study, the most optimal short form of the Token Test.

4.11 TEA Visual Elevator

4.11.1 Literature Review and Introduction

No relevant studies were obtained in the literature search.

The visual elevator (VE) is one of the eight subtests of the Test of Everyday Attention (TEA). As the name of the TEA implies, it is assumed that al subtests of the TEA are closely related to everyday tasks and therefore have a high ecological validity (Robertson, Ward, Ridgeway, & Nimmo-Smith, 1994). The VE is designed to assess attentional switching (Robertson et al., 1994; Robertson, Ward, Ridgeway & Nimmo-Smith, 1996; Robertson, Manly, Andrade, Baddeley & Yiend, 1997). Some studies however argue that not attentional switching but sustained attention is measured with this subtest (Bate et al., 2011).

In the TEA VE, patients have to imagine that they are in an elevator with a broken display, wherefore they do not know on which floor they will arrive. Using a printed sequence of arrows that indicate the (changing) direction of the elevator, patients have to decide on which floor they are. The TEA VE consists of ten sequences and two different scores can be conducted: the accuracy score (which measures the correctly counted items) and the timing score (the total time taken for correct items divided by the total number of switches, whereby a lower value indicates a better performance than a higher value). The timing score is able to discriminate between severe traumatic brain injury and healthy controls (Bate et al., 2001).

The administration time of the TEA VE varies a lot since the items can be easy for a patient or almost too difficult to complete. No cutoff point is defined, so if a patient is not capable of completing the items, the test leader should rely on his or her expertise to stop or continue the test. In addition, it is not possible to define a standard cut-off point prior to the assessment, because not only the time but also the accuracy is used to interpret the performance on the test. Therefore, another way to reduce the time of assessment will be evaluated in this study, namely that of shortening the test itself.

The literature search did not reveal any short forms. It did however reveal some information about the use of parallel forms. The test-retest agreement among three parallel forms of the TEA VE is evaluated in patients with chronic stroke (Chen, Kon, Hsieh & Hsueh, 2013). Reliability is good to excellent for the timing scores, but the accuracy scores had poor reliability for two of the three forms. Practice effects after an one week interval for patients with chronic stroke were medium when comparing the timing scores of parallel forms (Chen et al., 2013). The relatively high practice effect for the TEA VE can probably be explained by the timing component, since this factor is known to be sensitive to practice (Lezak et al., 2012). When developing a short form, this information should be taken into account.

Although the literature search did not reveal any information about short forms of the TEA VE, a short form of the TEA VE is developed in the Netherlands (in the University Medical Center of Utrecht) for practical use. In this short form, 50% of the original items are administered. There is deliberately chosen for these items, so that the number of switches in the short form is representative to the number of switches in the original test, both in terms of variation of the number of switches for each item as in terms of the total amount of switches for the items (which is exactly 50% of the original total amount of switches). Scores are defined as in the original form, so that they can be compared to the original norms when multiplied by two. In this thesis, this shortened version will be compared to the original TEA VE to evaluate if this short form is sufficient for the use in neuropsychological assessment.

4.11.2 Methods

The short form of the TEA VE, consisting of items 1, 4, 6, 7, and 10 of version A, will be compared to the original version A of the TEA VE. The reliability, equivalence, consistency between scores, and diagnostic agreement will be investigated as described in the general methods of this thesis (see paragraph 2.3). Note that for this test both the accuracy scores and the timing scores are determined. Therefore, the accuracy scores of the original and short version will be compared to each other, just as the timing scores of the original and short version. Timing scores are defined as the number of seconds that were necessary to complete an item divided by the number of switches of that item and are only determined for correct items. To evaluate if the relations of the two versions and an external construct are equivalent, the relations with the Brixton Spatial Anticipation Test (original test version) are considered.

When defining the diagnostic agreement, the norm scores provided by Robertson and colleagues (1994; appendix 4b and 4d) will be used to classify the scores as excellent, above average, average, below average, or impaired, in the way suggested by Lezak (2012; see paragraph 2.3). The same norms will be used to classify the scores of the shortened version, after these scores are multiplied by two.

4.11.3 Results

Reliability

The degree of consistency across items was assessed for both the original and the short version of the TEA VE. For the timing scores, this degree was very high in the original version, Cronbach's $\alpha = 0.95$, and high in the short form, Cronbach's $\alpha = 0.88$. On the other

hand, the accuracy had an adequate internal consistency in the original version, Cronbach's α = 0.70, and a low internal consistency in the short form, Cronbach's α = 0.24.

Equivalence

The relation between the short form and the original TEA VE is examined. Spearman's correlation coefficients were used, since the data for all scores are non-normally distributed. The relation between the accuracy scores of the two forms is strong (r_s = 0.80, p < 0.01), just as the relation between the timing scores of the two forms (r_s = 0.97, p < 0.01). All relations can be found in table 10.

The second index of equivalence, parallelism, is determined by comparing the relations between the original form and demographic variables (namely age, sex, and level of education) with the relations between the short form and these demographic variables. If the short form measures the same construct as the original form, it should have approximately the same correlations with these variables. Firstly, the Spearman correlation coefficients were determined. The timing scores were significantly correlated with level of education, whereby participants with an higher level of education perform better (r_s = -0.70, p < 0.01 for the original form; r_s = -0.69, p < 0.01 for the short form). Although the timing scores of the short form were significantly correlated with age (r_s = 0.44, p < 0.01), the relation between the timing scores of the original form and age could not be classified as significant (r_s = 0.35, p = 0.118). Secondly, the relations of the two test versions with these variables are compared, to see whether they are in the same range. As can be seen in the last two rows of table 11, the relations between the different demographic variables and the timing scores of the two test versions are not significantly different and are therefore in the same range, just as the relations with these variables and the accuracy scores of the two test versions. In other words, there is parallelism.

Table 10. *Spearman's correlation coefficients between the different scores of the original and short form.*

	VE Total: Accuracy score	**VE Total: Timing score**	**VE Short: Accuracy score**
VE Total: Timing score	-0.12		-0.20
VE Short: Accuracy score	0.80**	-0.20	
VE Short: Timing score:	-0.18	0.97**	-0.29

*** p < 0.01.*

Table 11. *Spearman's correlation coefficients between the total scores of the two forms and respectively age, sex, and education and the difference between these correlations.*

	Age	Sex	Education
VE Total: Timing score	r_s = 0.35	r_s = 0.08	r_s = -0.70**
VE Short: Timing Score	r_s = 0.44*	r_s = 0.13	r_s = -0.69**
VE Total: Accuracy score	r_s = -0.024	r_s = -0.067	r_s = 0.197
VE Short: Accuracy score	r_s = -0.081	r_s = 0.184	r_s = 0.113
Difference in correlations between the timing scores	z = 1.69 (*p* = 0.09)	z = 0.81 (*p* = 0.42)	z = 0.24 (*p* = 0.81)
Difference in correlations between the accuracy scores	z = 0.38 (*p* = 0.70)	z = 1.71 (*p* = 0.09)	z = 0.57 (*p* = 0.57)

* $p < 0.05$. ** $p < 0.01$.

The correlations of the two test versions with an external construct are also compared to each other, namely the correlations with the Brixton Spatial Anticipation Test (original test version). The Spearman's correlation coefficients are not significant for the timing scores (r_s = -0.11, p = 0.64 for the original test; r_s = -0.12, p = 0.60 for the short form) and for the accuracy scores (r_s = -0.03, p = 0.91 for the original test; r_s = 0.11, p = 0.64 for the short form). Moreover, these correlations do not differ significantly when comparing the two test versions (z = 0.17, p = 0.87 for the timing scores, z = 0.94, p = 0.35 for the accuracy scores). Therefore it is concluded that the short form is representative to the original form with regard to their relation with the Brixton Spatial Anticipation Test, an external construct.

Consistency between scores

The consistency between the ranking positions for the test scores are compared to each other using a related-samples Wilcoxon Signed Rank Test. The ranking variables for the total scores do not differ significantly when the accuracy scores of the two test versions are compared (p = 0.82), meaning that the ranks in the sample are assigned equally for both the original test and the short form. The same conclusion applies for the consistency of the timing scores in both test versions, since the ranking variables for these scores do not differ significantly either (p = 0.83).

Diagnostic Agreement

In the sample, the diagnostic categories 'below average', 'average', and 'above average' were assigned to the accuracy scores in both test versions. All diagnostic categories ('impaired', 'below average', 'average', 'above average' and 'excellent') were assigned to the

timing scores in both test versions. The rate of agreement between the original version of the TEA VE and the short form is almost perfect, both when the accuracy scores are considered, $\kappa = 0.867$, as when the timing scores are considered, $\kappa = 0.926$.

4.11.4 Conclusion and Discussion

In this study, a short form of the TEA VE with items 1, 4, 6, 7, and 10 is compared to the original TEA VE. The reliability of the original test has shown to similar to the reliability found in previous studies (e.g. Chen et al., 2013). The current results show that the reliability of the original test and that of the short form are both high. That is, when the timing scores are considered. For the accuracy scores, the reliability of the original test is only adequate and even low in the short form, wherefore the short form cannot be used to interpret the accuracy scores.

There is a strong relation between the short form and the original TEA VE, especially regarding the timing variables. Also, the relations between the short form and demographic variables and an external construct are in the same range as the relations between the original test and these variables, both when the timing scores as when the accuracy scores are considered. Moreover, for both the accuracy as the timing scores the ranks where assigned equally in the sample with both test versions and the diagnostic agreement is almost perfect. This indicates that the short form is equivalent to the original TEA VE and can be used for diagnostic purposes.

It is noteworthy that the psychometric quality is constantly rated better when the timing scores instead of the accuracy scores are considered. This is in line with earlier research (e.g. Chen et al., 2013), in which is stated that the timing score should be used as a measure of attention, and not the accuracy scores. The explanation for this is that participants are required to perform the task not only as accurate, but also as quickly as possible, whereby they have to find a balance between speed and accuracy. Since the timing score considers both speed and accuracy, this score can represent such an speed-accuracy tradeoff, while the accuracy score cannot (Chen et al., 2013). It is therefore recommended to interpret only the timing scores and not the accuracy scores when evaluating the performance in both the short version as in the original TEA VE. Hence, the low reliability of the accuracy scores in the short form do not have to be reviewed as a disadvantage.

Although general limitations of and recommendations based on this thesis can be found in paragraph 5, one important limitation of the current study about the short form of the TEA VE needs to be stressed. For this study, the rearrangement of items for the short form occurred post hoc. Because all items and not only the items of the short form were assessed when collecting the data, theoretically some bias, caused by fatigue or learning from

feedback or practice, could be present. This is especially important in this NP test, since practice effects are known to be relatively high for the TEA VE (Chen et al., 2013). In future studies only the short form should be administered, to control for this bias or to collect normative data that is certainly suitable for this short form.

In short, this study evaluates the first short form of the TEA VE that is known to be developed. When the timing scores and not the accuracy scores are considered, this short form is not only representative to the original test, but also suited for the use as a stand-alone measure for diagnostic purposes.

5. General Conclusion and Discussion

Efficient neuropsychological assessment is required because of several factors. In the first place, the frequency of neuropsychological evaluation (NPE) is increasing. Additionally, the current financial health climate demands cost-effective NPE. Further, a long assessment time not only has negative influences on the quality of the NPE due to factors such as fatigue or frustration but also on the wellbeing of a patient. Because screening instruments are not suitable for diagnostic purposes, this thesis focuses on the reduction of assessment time by shortening frequently used neuropsychological (NP) tests in order to achieve efficiency in neuropsychological assessment by reducing its duration.

Eleven frequently used NP tests were evaluated in this thesis. Note that these tests all evaluate different cognitive (sub)domains, so that efficiency within the NPE of all cognitive domains is considered. For all eleven tests, it is reviewed whether parallel short forms are developed. This was the case for seven tests. For the RAVLT, short forms with less than fifteen words were developed, but because these forms are not suitable for most patients a short version is reviewed with three instead of five trials of direct recall. Although the diagnostic agreement could not be examined in this thesis, the three-trial version seems to be highly equivalent to the five-trial version and may be even more advisable since ceiling effects are less likely to occur in the short form. The results concerning the BNT, for which the 29-items version was compared to a 15-items version, were not that positive. The quality of the 15-items version did not even meet the criteria for the use as a screening instrument. For both the phonological verbal fluency test as well as the semantic verbal fluency test it is suggested to use the longer version in NPE. For the JLO, many short forms were developed in previous studies. The evaluated 20-items short form is suitable in NPE, but the evaluated 15-items short form is insufficient for the use in NPE. The quality of the two half forms of the 21-items Token Test is not sufficient for the use in NPE either. The evaluated short form of the TEA VE can replace the original test in NPE. That is, when the timing scores and not the accuracy scores are considered.

For the other four selected NP tests, no short forms have been developed in previous studies. For the WAIS Digit Span, RCF, and Trail Making Test different options of abbreviation were discussed, but abbreviation did not seem beneficial or possible. For the RCF, one possibility of abbreviation (the development of another figure) could be examined in future research. In addition, in the WAIS Digit Span and the Token Test administration time is already restricted. For the BSAT no short forms have been developed either, but abbreviation did seem beneficial. Therefore, a short form is developed based on item

difficulty. This short form did however not have the same diagnostic quality as the original BSAT and should therefore not be used as a stand-alone measure in NPE.

Most previous studies on test abbreviation only review the reliability of the short forms and their relation with the original test. Only a few studies were found that investigate parallelism or the diagnostic agreement between the short form and the original test. In these studies, the diagnostic agreement is however not corrected for chance. In the current thesis, a combination of several measures was used to gain insight in the representativeness of the short forms, including a measure of diagnostic agreement which corrects for chance. It turned out that this measure was the most conservative and did not always reveal a good degree of diagnostic agreement, even though the other measures of equivalence confirmed parallelism between the short and the original test version. It is therefore of great importance to investigate the diagnostic agreement when short forms are evaluated. Consequently, short forms should only be used for diagnostic purposes in clinical settings when the diagnostic agreement with the original test is determined. This includes short forms that were developed in previous studies in which diagnostic agreement was not taken into account. In the current thesis, the diagnostic agreement could not be determined for all short forms and should therefore be assessed in future studies. Based on the findings in this thesis, it is strongly advised for future research on test abbreviation to consider not only the reliability and relation with the short and original test. Rather, more measures of equivalence should be considered, of which the diagnostic agreement is of great importance.

Despite the described advantage of the methods of this thesis, some limitations need to be stressed. First, it must be noted that, when evaluating the short forms, the arrangement of items happened post-hoc. That is, the full test versions were assessed when collecting data. Decline in performance due to fatigue or improvement in performance due to learning from feedback or practice could therefore have allowed some bias in the scores of the short forms. Future research should investigate whether changes in performance occur when the short form is presented beforehand. Second, the test-retest reliability was not assessed in this study. Previous studies showed that the test-retest reliability of short forms can vary from the original test (e.g. Harrison et al., 2000). Therefore, the test-retest reliability of the short forms that were pointed out as promising in the current thesis should be assessed in future research. At last, the current thesis uses data of healthy adults for the analyzes. Patients probably make more errors on several tests than healthy adults. Although it is reviewed whether the short forms are able to discriminate between the different diagnostic categories that were present in the current study, it is also important that the short forms are able to discriminate when the scores are lower, i.e. to discriminate between impaired and unimpaired scores. In addition, it is possible that patients not only make more errors during NPE, but also different kind of errors (e.g. repetitions in the RAVLT). The short forms should

be able to detect these different kind of errors equally to the original tests. For these two reasons, it is important to review the quality and diagnostic utility of the short forms within patient groups as well.

Besides the methodical limitations of this study, there are also disadvantages of the reduction of assessment time of NPE in general. One important limitation of a reduced assessment time is that neuropsychologists or its test assistants have less time to observe the patient. Observations are an important part of NPE. Not only emotional responses or coping strategies can be observed, but expressions of brain pathology can also be present. An example is the fluctuation of performance, which is part of some brain pathologies (for example in Lewy Body Dementia). When the time of NPE is reduced there is of course less time to observe such fluctuations, but the fluctuations are also less likely to occur during a shortened test than during a longer test. This increases the chance of misinterpretation of the test results. A second limitation of the use of abbreviated test versions is that the psychometric and diagnostic quality will always be slightly reduced. In clinical settings, the loss of psychometric quality and observation time due to the use of abbreviated test forms should always be considered and be compared to the advantages of time gain. Therefore, not only the choice of NP tests but also the choice of the test version (original or short) requires expert opinions of neuropsychologists.

Furthermore, it is known that neuropsychologists are conservative in their choice of tests. Many NP tests are used for over decennia. There are of course advantages of using the same tests for a long time, for example regarding the collection of normative data. However, it could be possible that, with the current knowledge within the neuropsychology, other NP tests can be developed that are equally good in evaluating a cognitive (sub)domain but are more efficient regarding their duration. Thus, the efficiency of NPE should be reviewed within a broader perspective.

Summarizing, in this thesis different short forms of frequently used NP tests have been evaluated. Multiple methods were used to examine their equivalence with the original test, of which diagnostic agreement has shown to be of great importance. Some short forms can be used in NPE for diagnostic purposes, but other short forms should not even be used as screening instruments. Although the use of short forms can definitely contribute to efficient neuropsychological assessment, their quality should be carefully examined before using the short forms in NPE. With this thesis, I hope to have boosted interest in this field of research.

6. References

Appels, B. A., & Scherder, E. (2010). Review: The Diagnostic Accuracy of Dementia-Screening Instruments With an Administration Time of 10 to 45 Minutes for Use in Secondary Care: A Systematic Review. *American Journal of Alzheimer's Desease and Other Dementias, 25,* 301-316.

Au, R., Joung, P., Nicholas, M., Obler, L.K., Kass, R., & Albert, M.L. (1995). Naming ability across the adult life span. *Aging and Cognition, 2,* 300–311.

Atkinson, T. M., Ryan, J. P., Kryza, M., & Charette, L. M. (2011). Using Versions of the Trail Making Test as Alternate Forms. *The Clinical Neuropsychologist, 25,* 1193-1206.

Baarda, D. B., Goede de, M. (2007). *Basisboek statistiek met SPSS*. Amsterdam: Noordhoff Uitgevers.

Babikian, T., Boone, K. B., Lu, P., & Arnold, G. (2006). Sensitivity and Specificity of Various Digit Span Scores in the Detection of Suspect Effort. *The Clinical Neuropsychologist, 20,* 145-159.

Bate, A. J., Mathias, J. L., & Crawford, J. R. (2001). Performance on the Test of Everyday Attention and Standard Tests of Attention following Severe Traumatic Brain Injury. *The Clinical Neuropsychologist, 15,* 405-422.

Benton, A.L., Hamsher, K., Varney, N. R. & Spreen, O. (1983). *Contributions to neuropsychological: A clinical manual.* New York: Oxford University Press.

Benton, A. L., & Hamsher, K. (1989). *Multilingual aphasia examination*. Iowa City: AJA.

Berg, E. van den., Nys, G. M. S., Brands, A. M. A., Ruis, C., Zandvoort, M. J. E. van., & Kessels, R. P. C. (2009). The Brixton Spatial Anticipation Test as a Test for Executive Function: Validity in Patient Groups and Norms for Older Adults. *Journal of International Neuropsychological Society, 15,* 695-703.

Blackburn, H .L., & Benton, A. L. (1957). Revised Administration and Scoring of the Digit Span Test. *Journal of Consulting Psychology, 21,* 139-143.

Bouma, A., Mulder, J. Lindeboom, J. & Schmand, B. (Eds.) (2012). *Handboek neuropsychologische diagnostiek.* Amsterdam: Pearson Assessment and Information B.V.

Brand, N. & Jolles, J. (1985). Learning and retrieval rate of words presented auditory and visually. *Journal of General Psychology, 112,* 201-220.

Brown, E. C., Casey, A., Fisch, R. I., & Neuringen, C. (1958). Trail Making Test as a Screening Device for the Detection of Brain Damage. *Journal of Consulting Psychology, 22,* 469-474.

Burgess, P. W., & Shallice, T. (1997). *The Hayling and Brixton Tests*. Thurston, UK: Thames Valley Test Company.

Butler, M, Retzlaaff, P., & Vanderploeg, R. (1991). Neuropsychological Test Usage. *Professional Psychology: Research and Practice, 22,* 510-512.

Calamia, M., Markon. K., Denburg, N. L., & Tranel, D. (2011). Developing a Short Form of Benton's Judgment of Line Orientation Test: An Item Response Theory Approach. *The Clinical Neuropsychologist, 25*, 670-684.

Canning, S. J. D., Leach, L., Stuss, D., Ngo, L. & Black, S. E. (2004). Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia. *Neurology, 62,* 556-562.

Chan, R. C. K., Shum, D., Toulopoulou, T., & Chen, E. Y. H. (2008). Assessment of Executive Functions: Review of Instruments and Identification of Critical Issues. *Archives of Clinical Neuropsychology, 23,* 201-216.

Chen, H. C., Koh, C. L., Hsieh, C. L., & Hsueh, I., P. (2013). Test of Everyday Attention in patients with chronic stroke: Test–retest reliability and practice effects*. Brain Injury, 27*, 1148-1154.

Cullen, B., O'Neill, B., Evans, J. J., Coen, R. F., & Lawlor, B. A. (2007). A review of screening tests for cognitive impairment. *Journal of Neurology, Neurosurgery & Psychiatry, 78,* 790-799.

De Renzi, E., & Vignolo, L. A. (1962) The Token Test: a sensitive test to detect receptive disturbances in aphasics*. Brain, 85,* 665-678.

Deelman, B., Eling, P., Haan de, E., & Zomeren van, E. (2009). *Klinische Neuropsychologie* (3rd ed.). Amsterdam: Boom.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California Verbal Learning Test Manual.* San Antonia, Texas: The Psychological Corporation.

Dongen van, H. R., Harskamp van, F., & Luteijn, F. (1976). *Tokentest: Handleiding.* Nijmegen: Berkhout bv.

Egberink, I.J.L., Janssen, N.A.M., & Vermeulen, C.S.M. (2009-2014). *COTAN Documentatie* (www.cotandocumentatie.nl). Amsterdam: Boom test uitgevers.

Elst, W. van der, Boxtel, M. P. J. van, Breukelen, G. J. O. van, & Jolles, J. (2005). Rey's verbal learning test: Normative data for 1855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society, 11,* 290-302*.*

Evers, A., Lucassen, W., Meijer, R., Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de Kwaliteit van Tests.* Amsterdam: Boom test uitgevers.

Fastenau, P. S., Denburg, N. L., & Mauer, B. A. (1998). Parallel Short Forms for the Boston Naming Test: Psychometric Properties and Norms for Older Adults. *Journal of Clinical and Experimental Neuropsychology, 20*, 828-834.

Field, A. (2009). *Discovering Statistics Using SPSS*. London: Sage.

Fillenbaum, G. G., Huber, M. H., & Taussig, I. M. (1997). Performance of eldery white and African American community residents on the abbreviated CERAD Boston naming test. Journal of Clinical and Experimental Neuropsychology, 19, 204-2010.

Geffen, G. M., Butterworth, P., & Geffen, L. B. Test-Retest Reliability of a New Form of the Auditory Verbal Learning Test (AVLT). *Archives of Clinical Neuropsychology, 9,* 303-316.

Goodwin, C. J. (2008). *Research in psychology: Methods and design* (5th ed.). New York: John Wiley & Sons.

Graves, R. E., Bezeau, S. C., Fogarty, J., & Blair, R. (2004). Boston Naming Test Short Forms: A Comparison of Previous Forms with New Item Respons Theory Based Forms. *Journal of Clinical and Experimental Neuropsychology, 26,* 891-902.

Gullet, J. M., Price, C. C., Nguyen, P., Okun, M. S., Bauer, R. M., & Bowers, D. (2013). Reliability of Three Benton Judgment of Line Orientation Short Forms in Idiopathic Parkinson's Disease. *The Clinical Neuropsychologist, 27,* 1167-1178.

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 Clinnical Psychology, 39,* 181-191.

Hendriks, M., Kessels, R., Gorissen, M., Schmand, B. (2010). *Neuropsychologische diagnostiek. De klinische praktijk* (2nd ed.). Amsterdam: Boom

Ingles, J. L., Eskes, G. A., Phillips, S. J. (1990). Fatigue after stroke. *Physical Medicine and Rehabilitation, 80,* 173-178.

Jonker, C., & Lindeboom, J. (1989). *Amsterdamse Dementie-Screeningtest*. Pearson: Amsterdam.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *Boston Naming Test.* Philadelphia: Lea & Febiger.

Karlsen, K., Larsen, J. P., Tandberg, E., & Jørgensen. (1999). Fatigue in patients with Parkison's Disease. Movement Disorders, 14, 237-241.

Kinsinger, S. W., Lattie, E., Mohr. D. C. (2010). Relationship between depression, fatigue, subjective cognitive impairment, and objective neuropsychological functioning in patients with multople sclerosis. Neuropsychology, 24, 573-580.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Lansing, A. E., Ivnik, R. J., Cullum, C. M. & Randolph, C. (1999). An Empirically Derived Short Form of the Boston Naming Test. *Archives of Clinical Neuropsychology, 14,* 481-487.

Lee, I. A., & Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from http://quantpsy.org.

Loon – Vervoorn, W.A. van, Stumpel, H.J., de Vries, L.A. (1995). *De Boston Benoemings-taak. Een test voor woordvinding bij afasie.* Utrecht.

Loon – Vervoorn, W. A. van. (2005, vierde druk). *De Boston Benoemingstaak. Een test voor woordvinding bij afasie. Normering voor Nederland.* Utrecht.

Lundin, K. A., & DeFilippis, N. A. (1999). Proposed Schedule of Usual and Customary Test Administration Times. *The Clinical Neuropsychologist, 13,* 433-436.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological Assessment* (5th ed.). New York: Oxford University Press.

Meyers, J. E., Zellinger, M. M., Kockler. T., Wagner, M., & Miller, R. M. (2013). A Validated Seven-Subtest Short Form for the WAIS- IV. *Applied Neuropsychology, 20*, 249-256.

Morris, J.C., Heyman, A., Mohs, R.C., Hughes, J.P., van Belle, G., Fillenbaum, G. et al. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I: 165. Clinical and Neuropsychological assessment of Alzheimer's disease. *Neurology, 39,* 1159-1165.

Mount, D. L., Hogg, J., & Johnstone, B. (2002). Applicability of the 15-item versions of the Judgment of Line Orientation Test for Individuals with Traumatic Brain Injury. *Brain Injury, 16,* 1051-1055.

Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a Brief Screening Tool for Mild Cognitive Impairment. Journal of the American Geriatrics Society, 53, 695-699.

Nunally, J., & Bernstein, I. H. (1994). *Psychometric theory (3rd ed.).* New York: McGraw Hill.

Osterrieth, P.A. (1944). Le test de copie d'une figure complex: Contribution à l'étude de la perception et de la memoire [The test of copying a complex figure: A contribution to the study of perception and memory]. *Archives de Psychologie*, *30*, 286-350.

Qualls, C. E., Bliwise. N. G., & Stringer, A. Y. (2000). Short Forms of The Benton Judgment of Line Orientation Test: Development and Psychometric Properties. *Archives of Clinical Neuropsychology, 15,* 159-163.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment Practices of Clinical Neuropsychologists in the United States and Canada: A Survey of INS, NAN, and APA Division 40 Members. *Archives of Clinical Neuropsychology, 20,* 33-65.

Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique [Psychological examination of traumatic encephalopathy] *Archives de Psychologie*, *28*, 286-340.

Rey, A. (1964). *L'examin Clinique en psychologie.* Paris, France: Presses Universitaires de France.

Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1994). *The Test of Everyday Attention*. Flempton: Thames Valley Test Company.

Robertson, I.H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1996). The structure of normal human attention: The test of everyday attention. *Journal of the International Neuropsychological Society, 2,* 525-534.

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). Oops: Performance Correlates of Everyday Attentional Failiures in Traumatic Brain Injured and Normal Subjects. *Neuropsychologia, 24,* 636-647.

Ruff, R. M., Light, R. H., & Parker, S. B. (1996). Benton Controlled Oral Word Association Test: Reliability and Updated Norms. *Archives of Clinical Neuropsychology, 11,* 329-338.

Saxton, J., Ratcliff, G., Munro, C. A., Coffey, E. C., Becker, J. T., Friend, L. & Kuller, L. (2010). Normative Data on the Boston Naming Test and Two Equivalent 30-Item Short Forms. *The Clinical Neuropsychologist, 14*, 526-534.

Schilder, C., Dijk, S. van, Meinhardt, W., Dam, F. van, Schagen, S. (2011). Cognitief functioneren van prostaatkankerpatiënten die hormonale therapie ondergaan. *Neuropraxis, 1*, 20-26.

Shin, M. S., Park, S. Y., Park, S. R., Seo, S. H., & Kwon. J. S. (2006). Clinical and empirical applicatins of the Rey-Osterrieth Complex Figure Test. *Nature Protocols, 1,* 892-899.

Spencer, R. J., Carrington, R. W., Giggey, P. P., Seliger, S. T., Katze, L. I., & Waldstein, S. R. (2013). Judgment of Line Orientation: An Examination of Eight Short Forms. *Journal of Clinical and Experimental Neuropsychology, 35,* 160-166.

Spreen, O., & Strauss, E. (1991). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*. New York: Oxford University Press.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245-251.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary.* New York: Oxford University Press.

Strauss, E. & Spreen, O. (1990). A comparison of the Rey and Taylor figures. Archives of Clinical Neuropsychology, 5, 417-420.

Thurstone, L. L. (1938). *Primary Mental Abilities.* Chicago: Chicago University Press.

Uchiyama, C. L., D'Elia, L. F., Dellinger, A. M., Beckert, J. T., Selnes, O. A., Wesch, J. E. Chen, B. B., Sats, P., Gorp, W. van, Miller. E. N. (1995). Alternate Forms of the Auditory-Verbal Learning Test: Issues of Test Comparability, Longitudinal Reliability, and Moderating Demographic Variables. *Archives of Clinical Neuropsychology, 10*, 133-135.

Uttner, I., Wittig, S., Arnim von, C. A. F., & Jäger, M. (2013). Kurz un einfach ist nicht immer besser: Grenzen kognitiver Demenzscreenings. *Fortschritte der Neurologie Psychiatrie, 81,* 188-194.

Vanderploeg, R. D., LaLone, L. V., Greblo, P., & Schinka, J. A. (1997). Odd-Even Short Forms of the Judgment of Line Orientation Test. *Applied Neuropsychology, 4,* 244-246.

Verhage, F. (1964). *Intelligentie en leeftijd.* Assen: Van Gorcum.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale Fourth Edition.* San Antonio, TX: Pearson Assessment.

Winegarden, B. J., Yates, B. L., Moses, J. A., Benton A. L., & Faustman, W. O. (1998). Development of an Optimally Reliable Short Form for Judgment of Line Orientation. *The Clinical Neuropsychologist*, 12, 311-314.

Woodard, J. L., Benedict, R. H. B., Roberts, V. J., Goldstein, F. C., Kinner, K. M., Capruso, D. X., & Clark, A. N. (1996). Short-form Alternatives to the Judgment of Line Orientation Test. *Journal of Clinical and Experimental Neuropsychology, 18,* 898-904.

Woodard, J. L., Benedict, R. H. B., Salthouse, T. A., Toth, J. P., Zgaljardic, D. J. & Hancock, H. E. (1998). Normative Data for Equivalent, Parallel Forms of the Judgment of Line Orientation Test. *Journal of Clinical and Experimental Neuropsychology, 20*, 457-462.

Wymer, J. H., Rayls, K., & Wagner, M. T. (2003). Utility of a Clinically Derived Abbreviated Form of the WAIS- III. *Archives of Clinical Neuropsychology, 18,* 917-927.

Zakzanis, K. K., McDonald, K., & Troyer, A. K. (2013). Component analysis of verbal fluency scores in severe traumatic brain injury. *Brain Injury, 27,* 903-908.

Zhao, Q., Lv, Y., Zhou, Y., Hong. Z., & Guo, Q. (2012). Short-term Delayed Recall of Auditory Verbal Learning Test is Equivalent to Long-Term Delayed Recall for Identifying Amnestic Mild Cognitive Impairment. *Plos one, 7,* e51157.

Zino, C., Ponsford, J. (2006). Selective attention deficits and subjective fatigue following traumatic brain injury. Neuropsychology, 20, 383-390.

Appendix 1. Summary of all analyzes

Table 12. *Summary of all analyzes, in which the criteria described in paragraph 2.3 are used to classify the results as sufficient (+) or insufficient (-).*

Test version	Inter-item reliability	Relation with long form	Difference in relation with age	Difference in relation with sex	Difference in relation with education	Difference in relation with an external construct	Consistency between ranking positions	Diagnostic Agreement
RAVLT trials 1-3 (versus trials 1-5)		+	+	+	+	+	+	
Boston 15 items (versus 29 items)	-	+	-	+	+	+	+	-
Phonological Fluency Test letter N (versus letters N&A)		+	+	+	+	+	+	-
Phonological Fluency Test letter A (versus letters N&A)		+	+	+	+	+	+	-
Semantic Fluency Test one minute (animals) (versus two minutes)		+	+	-	+	+	+	
BSAT 5 series (versus 9 series)	+	+	+	+	+	+	+	-
JLO V10-24 (versus V1-30)	-	+	+	+	+	+	+	-
JLO V11-30 (versus V1-30)	+	+	+	+	+	+	+	+
Token Test even items (versus 21 items)	-	+	+	+	+	+	+	-
Token Test uneven items (versus 21 items)	-	+	+	+	+	+	+	-
TEA VE 5 items timing (versus 10)	+	+	+	+	+	+	+	+
TEA VE 5 items accuracy (versus 10)	-	+	+	+	+	+	+	+

Appendix 2. The 29-item version of the Boston Naming Task[1]

1. Helikopter	**Helicopter**
2. Inktvis	**Octopus**
3. Paddestoel	Mushroom
4. Masker	Mask
5. Vulkaan	**Volcano**
6. Zeepaard	**Seahorse**
7. Bever	Beaver
8. Mondharmonica	Harmonica
9. Neushoorn	**Rhinoceros**
10. Eikel	**Acorn**
11. Domino	Dominoes
12. Cactus	Cactus
13. Harp	**Harp**
14. Klopper	Knocker
15. Pelikaan	**Pelican**
16. Stethoscoop	Stethoscope
17. Piramide	**Pyramid**
18. Muilkorf	Muzzle
19. Eenhoorn	Unicorn
20. Trechter	Funnel
21. Strop	Noose
22. Passer	Compass
23. Statief	**Tripod**
24. Oorkonde	**Muniment**
25. Sfinx	**Sphinx**
26. Juk	Yoke
27. Palet	**Palette**
28. Gradenboog	**Protractor**
29. Telraam	**Abacus**

Note. Items on the 15-item version are in bold face.

[1] Kaplan, Goodglass & Weintraub (1983); Van Loon – Vervoorn (2005).

Appendix 3. The 21-item version of the Token Test[2]

1	Leg de rode cirkel op de groene rechthoek. *Put the red circle on the green circle.*
2	Leg de witte rechthoek achter de gele cirkel. *Put the white square behind the yellow circle.*
3	Raak de blauwe cirkel met de rode rechthoek aan. *Touch the blue circle with the red square.*
4	Raak met de blauwe cirkel de rode rechthoek aan. *Touch the blue circle with the red square.*
5	Raak de blauwe cirkel of de rode rechthoek aan. *Touch the blue circle or the red square.*
6	Raak de blauwe cirkel en de rode rechthoek aan. *Touch the blue circle and the red square.*
7	Haal de groene rechthoek bij de gele rechthoek weg. *Remove the green square from the yellow square.*
8	Leg de rode cirkel voor de groene rechthoek. *Put the red circle before the green square.*
9	Als er een zwarte cirkel is, pak dan de rode rechthoek op. *If there is a black circle, take the red square.*
10	Pak de rechthoeken op behalve de gele. *Take the squares except the yellow one.*
11	Wanneer ik de groene cirkel aanraak, met u de witte rechthoek pakken. *When I touch the green circle, you should take the white square.*
12	Leg de groene rechthoek naast de rode cirkel. *Put the green square beside the red circle.*
13	De rechthoeken moet u langzaam na elkaar aanraken, de cirkels moet u snel na elkaar aanraken. *You should touch the squares slowly after each other, you should touch the circles quickly after each other.*
14	Leg de rode cirkel tussen de gele rechthoek en de groene rechthoek. *Put the red circle between the yellow square and the green square.*
15	Raak alle cirkels behalve de groene aan. *Touch all circles except the green one.*

[2] De Renzi & Vignolo (1962); van Dongen, van Harskamp & Luteijn (1976).

16	Pak de rode cirkel -neen- de witte rechthoek op. *Take the red circle -no- the white square.*
17	Pak de gele cirkel in plaats van de witte rechthoek. *Take the yellow circle instead of the white square.*
18	Pak de gele cirkel en de blauwe cirkel op. Take the yellow circle and the blue circle.
19	Nadat u de groene rechthoek heeft gepakt, moet u de witte cirkel aanraken. *After you have got the green square, you should touch the white circle.*
20	Leg de blauwe cirkel onder de witte rechthoek. *Put the blue circle below the white square.*
21	Voordat u de gele cirkel aanraakt, moet u de rode rechthoek pakken. *Before you touch the yellow circle, you should take the red square.*