Statistical Advances in Clinical Neuropsychology

Joost Agelink van Rentergem

COLOPHON

ISBN: 978-94-028-0961-9

Printed by Ipskamp Printing, Enschede

This thesis was typeset using LaTeX, and `classicthesis`, developed by André Miede and Ivo Pletikosić.

The style was inspired by Robert Bringhurst's book on typography "*The Elements of Typographic Style*".

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC® C128610

STATISTICAL ADVANCES IN CLINICAL NEUROPSYCHOLOGY

ACADEMISCH PROEFSCHRIFT
ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr.ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 27 maart 2018, te 12:00 uur
door Johannes Albert Agelink van Rentergem Zandvliet
geboren te Amsterdam

Promotiecommissie:

Promotores:	prof.dr. H.M. Huizenga, Universiteit van Amsterdam
	prof.dr. B.A. Schmand, Universiteit van Amsterdam
Copromotor:	prof.dr. J.M.J. Murre, Universiteit van Amsterdam
Overige leden:	prof.dr. C.V. Dolan, Vrije Universiteit Amsterdam
	prof.dr. H.M. Geurts, Universiteit van Amsterdam
	prof.dr. R.P.C. Kessels, Radboud Universiteit Nijmegen
	prof.dr. F.J. Oort, Universiteit van Amsterdam
	prof.dr. S.B. Schagen, Universiteit van Amsterdam
	prof.dr. M.E. Timmerman, Rijksuniversiteit Groningen

Faculteit Maatschappij- en Gedragswetenschappen

CONTENTS

1 GENERAL INTRODUCTION

Consider the case of an elderly man, Albert, who has fallen off his bicycle because of a reckless scooter driver, and is taken to hospital. Albert's physical injuries do not seem to be severe, and he is soon released. After some weeks, Albert's wife contacts the hospital, because she thinks his memory has worsened. Albert himself is not aware of any changes. A neurologist sends Albert to a clinical neuropsychologist for a neuropsychological assessment. The clinical neuropsychologist has to decide between several options. Is Albert's memory indeed bad, and is this consistent with a traumatic brain injury from his accident? Is Albert's wife perhaps overly worried, and is Albert's memory consistent with what would be expected for a man of his age? Or is Albert's memory bad, and is this part of a larger problem, perhaps a disorder like Alzheimer's disease?

It is important to Albert, his wife, and the hospital that the neuropsychological assessment is as reliable as possible. If Albert's wife is indeed overly worried, this should be discovered, so these worries can be resolved. If Albert is suffering from a traumatic brain injury, this should be discovered so the hospital can further investigate this injury (Maas, Stocchetti, & Bullock, 2008). If Albert is suffering from a disorder like Alzheimer's disease, this should be discovered so Albert can start with treatment (Small et al., 1997). It is therefore crucial that the neuropsychological assessment is successful in providing clarity to all parties.

Outside clinical practice, neuropsychological assessments are also performed in research. This may be done in studies that evaluate a new treatment, for example, to ameliorate the symptoms of dementia. Neuropsychological assessments are also used to detect adverse effects of treatments on cognitive functioning. Cognitive functioning may be affected by a wide variety of pharmaceutical treatments, for example psychiatric drugs (Moore & O'Keeffe, 1999) or drugs that are aimed at a different target entirely, like chemotherapy (de Ruiter et al., 2011), and non-pharmaceutical treatments such as deep brain stimulation (Smeding, Speelman, Huizenga, Schuurman, & Schmand, 2006), or surgery of the brain (Spencer & Huh, 2008) or heart (Selnes et al., 2012). In studying these treatments, it is important that the neuropsychological assessment is highly reliable, because otherwise, harmful side effects may be overlooked.

Studies may also use neuropsychological assessments to evaluate whether cognitive functions are impaired in a particular disorder, be it a disorder of the brain like schizophrenia (Schaefer, Giangrande, Weinberger, & Dickinson, 2013), or a disorder that may indirectly affect cognition, like liver cirrhosis (O'Carroll et al., 1991) or diabetes (Cheng, Huang, Deng, & Wang, 2013). If this is the case, researchers may study what characteristics of the patients predict which patients are affected, as some cognitive problems for example primarily occur in older patients. The reliability of the neuropsychological assessment is again critical in identifying those with cognitive impairment, and those without.

The goal of this thesis is to improve the reliability of neuropsychological assessment, specifically by improving the normative comparison procedure. This thesis is embedded in the Advanced Neuropsychological Diagnostics Infrastructure (ANDI) project. This thesis discusses multiple statistical methods that were developed for the ANDI project to improve normative comparisons. In this chapter, several key concepts are introduced, and the specific goals for the ANDI project are outlined. Then, an overview of the remaining chapters is given.

1.1 WHAT IS NEUROPSYCHOLOGICAL ASSESSMENT?

In clinical neuropsychology, patients are assessed to characterize their cognitive functioning. Subjective cognitive complaints, an accident or stroke, or a disorder like Parkinson's disease are all indications that cognitive functioning may be impaired, and can thus be reasons for a neuropsychological assessment (Lezak, Howieson, Bigler, & Tranel, 2012). This type of assessment is standardized, in order to make the results comparable between different clinicians and patients. Therefore, standardized neuropsychological tasks are used, which may consist of memorizing a message, naming objects in pictures, enumerating as many words starting with a particular letter as possible within one minute, or tracing a pattern with a pencil (Strauss, Sherman, & Spreen, 2006). Each of these tests is designed to tap into a different part of cognitive functioning, such as memory, psychomotor skills or attention. By measuring these cognitive functions, the neuropsychologist can decide whether a patient's cognition is impaired. The goal of the ANDI project and this thesis is to improve the precision with which this decision is made.

1.1.1 *Normative comparisons*

To decide whether a particular score on a test is indicative of impairment, a certain reference standard has to be used. For almost all neuropsychological tasks, there is no score that can be considered indicative of impairment in an absolute sense. Rather, patients' test scores are considered relative to those obtained by a group of healthy people (Crawford & Garthwaite, 2002), typically called a norm group or normative sample. If a patient's test score is lower than those obtained by the majority of healthy people, this is an indication for impairment. To be able to make such a judgment, called a normative comparison, data from many healthy participants who have completed neuropsychological tests are needed. Therefore, one of the goals of the ANDI project is to establish a large normative dataset, to improve normative comparisons.

1.1.2 *Multivariate normative comparisons*

The idea for the ANDI project came in part from the introduction of a new statistical technique for normative comparisons, called multivariate normative comparisons. Traditionally, normative comparisons are performed for a single neuropsychological test at a time, and are therefore univariate (Huizenga, Smeding, Grasman, & Schmand, 2007). This univariate approach has two disadvantages. The first is that it does not match clinical intuition, as results on tests are not interpreted in isolation by clinicians, but are interpreted in the light of results on other tests. For example, a low score on two delayed memory tests is interpreted differently when found in a patient with high scores on all other tests, than when found in a patient with low scores on all tests.

The second disadvantage is an increased number of times that a patient is incorrectly classified as cognitively impaired, i.e., that the assessment indicates impairment, while the patient is in fact not cognitively impaired. The aim is to keep the number of persons that are mislabeled like this low. However, for each normative comparison, there is a probability that this comparison will by chance indicate impairment, which is called a false positive result. This is the case even for a cognitively healthy person. With univariate normative comparisons, a comparison is performed for every neuropsychological test score, and the probability of at least one false positive result for a healthy person becomes larger and larger by chance when additional normative comparisons continue to be made. For example, a healthy person has a higher chance of a false positive result if this person is given many opportunities, in tests of verbal memory, executive functions, motor speed, attention, naming, and fluency. This risk is lower if only a single test is administered. There is no good way of knowing for a new patient whether a finding of cognitive impairment is incorrect or not, and if no steps are taken to control the number of times that incorrect classifications are made, many healthy people may inadvertently be labeled as cognitively impaired by univariate comparisons (Binder, Iverson, & Brooks, 2009).

These disadvantages are not found in multivariate normative comparisons. First, multivariate normative comparisons analyze the entire profile of test scores, similarly to how a clinician takes into account the whole profile of scores (Huizenga et al., 2007). This means that the analysis takes into account whether the patient's profile of scores is common in healthy participants. For example, the combination of a very high score on immediate recall of words and a low score on an attention test is common among healthy people. The combination of a very high score on immediate recall of words and a low score on recall of words after 30 minutes is something that is not observed in healthy people. This profile of scores could indicate impairment of memory storage.

Second, multivariate normative comparisons always provide a single comparison. This means that if a profile of twenty-five test scores is tested in a normative comparison, this entails a single comparison, just like a profile of five test scores would. Because there is only a single comparison, the probability of finding a false positive result, and thus incorrectly classifying a cognitively healthy person as cognitively impaired, is under control, no matter how many neuropsychological test scores are entered into the comparison.

One problem for multivariate normative comparisons is that it requires that the healthy people in the normative group have completed multiple tests. Ideally, they would have completed all the same neuropsychological tests that the patient completes in the assessment. This type of normative data is not available. Therefore, one of the goals of the ANDI project and this thesis is to provide normative data from healthy participants who have completed multiple tests, in order to facilitate the implementation of multivariate normative comparisons.

1.1.3 *Demographic corrections*

Another important aspect where normative comparisons can be improved is the area of demographic corrections. When evaluating a patient's scores for the presence of a cognitive impairment due to a disorder, the cognitive impairment is best detected when the healthy participants in the normative group are similar to the patient in characteristics unrelated to the disorder. What this means is that a neuropsychological assessment for a 72-year-old patient is most reliable when we compare his or her scores to those obtained by healthy 72-year-olds. Such corrections are commonly performed for age. However, level of education also predicts cognitive test scores. Therefore, we ideally compare test scores from patients with low education to those obtained by healthy people with low education, to increase sensitivity. Sex generally plays a smaller role, but there may be a small increase in sensitivity if male patients are compared to healthy men, and female patients are compared to healthy women (Lezak et al., 2012). Which demographic variables to correct for depends on the type of test used (de Vent, Agelink van Rentergem, Murre, ANDI Consortium, & Huizenga, 2016a, this thesis).

The available normative data for neuropsychological tests rarely allow demographic correction for age, sex, and level of education. Instead, normative data may be available for different ages, but not for different levels of education and sexes. Also, because demographic corrections require many different participants, data may not be available for individual ages. This means that, for example, a 72-year-old patient has to be compared to a group of 70 to 80-year-olds, which decreases sensitivity (Testa, Winicki, Pearlson, Gordon, & Schretlen, 2009). Therefore, one of the goals of the ANDI project and this thesis is to provide normative data from large numbers of healthy participants who have completed neuropsychological tests and for whom age, sex, and level of education is known, in order to facilitate more precise demographic corrections.

1.1.4 *Online availability*

A third major theme of this thesis and the ANDI project is using internet-based technology to aid clinical neuropsychology.

Normative comparisons for a single test that are corrected for age are typically performed by looking up the patient's score in a printed table of age bins and scores. Scores for different ages, sexes, and levels of education become more difficult to tabulate and to look up. The same is true for multivariate normative comparisons: Multivariate normative comparisons cannot be easily performed with printed tables, as there are many dimensions if there are multiple tests involved. One solution is to no longer look up the results by hand, but to let computers calculate the results (Miller & Barr, 2017). Therefore, one goal of the ANDI project is to build a website on which clinicians can perform normative comparisons of their patient data. This allows clinicians to use these statistically sophisticated techniques anywhere, and allows us to update the procedures as new data and methods become available.

Another advantage of using the internet is that it becomes easy to share information with a large number of clinicians and scientists. The normative comparison procedures described in this thesis are in principle not restricted to the field of clinical neuropsychology. Therefore, one could take the software and apply it in other fields of psychology or in other disciplines, such as medicine. To facilitate this, the computer code for the methods developed in this thesis and that are used in the ANDI project are freely available online. A second advantage of sharing the code of the ANDI project is transparency (Poldrack & Gorgolewski, 2014). This means that the implementation of the methods described in this thesis are also available to any user or programmer who wants to review and criticize the method (Nosek et al., 2015).

1.2 OVERVIEW OF THIS THESIS

In the second chapter, a method is described for establishing a normative dataset that fits the goals outlined above. This method is based on the combination of data from healthy people who have already taken part in research, for example as a participant in a control group in a clinical study, or as a participant in an epidemiological community study. By combining data from multiple studies, it becomes possible to obtain large numbers of participants, who are demographically diverse, and have completed many different tests. This chapter explains standardized procedures for removing outlying values, determining what demographic variables to use in corrections, and finding appropriate transformations that facilitate normative comparisons. Also, this chapter describes how these methods have been applied to data that were generously donated by the ANDI consortium, to form the ANDI database. A description of the contents of the ANDI database is also given.

In the third chapter, multivariate normative comparisons are described, and are extended to include demographic corrections. Also, it is explained how an aggregate database like the one described in chapter two can be used for normative comparisons. An aggregate database is different from standard normative datasets in that there may be differences between contributing studies in how participants perform. In this chapter, a multivariate multilevel regression model is introduced that resolves this issue. A second advantage of this model is that it can be fitted even when many data are missing. Missing data are very common in aggregate data, as some test variables may be completely absent from a particular study. In a simulation study, the appropriateness of the multivariate multilevel regression method is demonstrated. With this method, multivariate normative comparisons with demographic corrections can be made for the most common tests. Another issue related to missing data in aggregate databases, i.e. missing overlap, is left unsolved in this chapter. This issue is addressed in the next chapter.

In the fourth chapter, the method of the previous chapter is extended to solve the issue of missing overlap. If there are two tests that have not been administered together in any of the studies, there is no overlap between the two variables, and it becomes difficult to estimate a multivariate model. This situation would arise with tests that are less commonly administered, as it is more likely that these tests have not been administered together in any of the studies that are included in the aggregate database. Therefore, this prevents the inclusion of less commonly administered tests in the multivariate normative comparison. Two solutions are tested in this chapter, using either a multiple imputation or a factor model approach. This chapter ends with the recommendation that the problem of missing overlap can best be resolved using a factor model, but only if the factor model is an appropriate description for the included tests. Therefore, the goal of the next chapter is to find an appropriate factor model.

In the fifth chapter, different factor models for neuropsychological tests are compared. These models have been formulated in the literature, and make different distinctions in which test variables measure the same cognitive function. Some models are complex and contain many different cognitive functions, while others are simpler. In this chapter, a factor meta-analysis (Cheung & Chan, 2005) is performed. In this analysis, factor models are fitted to a correlation matrix that is pooled across multiple studies conducted worldwide. From this analysis, a single best fitting model is identified. Next, factor models are fitted to data from the ANDI database. Again, the best fitting factor model is identified. Together with the method described in the fourth chapter, this factor model allows for multivariate normative comparisons with more tests than was possible with the method from the third chapter.

In the sixth chapter, multivariate normative comparisons using ANDI are applied to address a clinical research question. The goal of this study is to classify patients with Parkinson's disease as either cognitively impaired or not, since impairment at an early stage of the disease is known to predict later development of Parkinson's disease dementia. With follow-up data that were gathered after three and five years, the performance in the prediction of dementia of the normative comparison procedure described in this thesis is compared to the performance of previously used methods. This thus provides an empirical test of the methods developed in this thesis.

In the seventh chapter, univariate normative comparisons using an aggregate database are discussed. As mentioned before, univariate comparisons can lead to incorrect classifications of cognitive impairment when many different comparisons are performed for different test variables. Therefore, if there is a scenario in which individual test scores are of interest rather than profiles of scores, there needs to be some kind of correction for false positives. In this chapter, several corrections that are described in the literature are discussed, and it is shown how they might be applied with an aggregate normative database. A new method is developed especially for this purpose.

In the eighth chapter, results of the previous chapters are summarized, and limitations and potential solutions are discussed. Possible extensions of the methods and the database, and possible applications outside the current scope of the ANDI project are discussed. The thesis ends with a consideration of how the ANDI project relates to recent developments in psychology.

2

ADVANCED NEUROPSYCHOLOGICAL DIAGNOSTICS INFRASTRUCTURE (ANDI): A NORMATIVE DATABASE CREATED FROM CONTROL DATASETS

2.1 ABSTRACT

In the Advanced Neuropsychological Diagnostics Infrastructure (ANDI), datasets of several research groups are combined into a single database, containing scores on neuropsychological tests from healthy participants. For most popular neuropsychological tests the quantity and range of these data surpasses that of traditional normative data, thereby enabling more accurate neuropsychological assessment. Because of the unique structure of the database, it facilitates normative comparison methods that were not feasible before, in particular those in which entire profiles of scores are evaluated. In this article, we describe the steps that were necessary to combine the separate datasets into a single database. These steps involve matching variables from multiple datasets, removing outlying values, determining the influence of demographic variables, and finding appropriate transformations to normality. Also, a brief description of the current contents of the ANDI database is given.

2.2 INTRODUCTION

An important element of neuropsychological practice is to determine whether a patient who presents with cognitive complaints has abnormal scores on neuropsychological tests. In the diagnostic process, a number of neuropsychological tests are administered and the test results of the patient are compared to a normative sample, that is, a group of healthy individuals which resemble the patient in characteristics unrelated to the suspected disease or trauma. In this manner, a clinician can determine whether the patient's test scores should be interpreted as abnormal, and whether or not the patient may have a disorder.

Traditionally, scores are compared to normative data published in the manuals of the neuropsychological tests. However, there are a number of limitations associated with this approach. First, normative

[0] Published as: de Vent, N. R.*, Agelink van Rentergem, J. A.*, Schmand, B. A., Murre, J. M. J., ANDI Consortium & Huizenga, H. M. (2016). Advanced Neuropsychological Diagnostics Infrastructure (ANDI): A normative database created from control datasets. *Frontiers in Psychology*, 7(1601), 1-10.

data of neuropsychological tests might have become outdated and no longer represent the patients we see today (Strauss et al., 2006). Second, many published tests lack norms for the very old population (80+; Whittle et al., 2007}. Third, some tests do not come with norms at all, and clinicians have to make do with norms from other countries or with norms they themselves have gathered (Crawford & Garthwaite, 2002). Fourth, normative scores from test manuals are often only corrected for age but not for other demographic variables, such as level of education or sex (Lezak et al., 2012). Fifth, normative data are typically collected for one test at a time, as part of its construction and standardization process. As a result, mostly univariate but not multivariate data are available. Recent studies have shown that multivariate normative comparison methods are more sensitive to deviating profiles of test scores than multiple univariate analyses (Crawford & Garthwaite, 2002; Huizenga et al., 2007; Smeding, Speelman, Huizenga, Schuurman, & Schmand, 2011; Schmand, de Bruin, de Gans, & van de Beek, 2010; Castelli et al., 2010; Valdés-Sosa et al., 2011; González-Redondo et al., 2012; Broeders et al., 2013; Cohen et al., 2014; Su et al., 2015). Moreover, new univariate methods for normative comparisons, that use a resampling technique, require multivariate normative data as well (Huizenga, Agelink van Rentergem, Grasman, Muslimovic, & Schmand, 2016).

Because of the limitations outlined above, we started the Advanced Neuropsychological Diagnostic Infrastructure project (www.andi.nl[1]). Our goal was to overcome these limitations by creating a large database from a demographically diverse group of healthy participants who completed several neuropsychological tests. This database will be accompanied by an interactive website where clinicians and researchers can upload their patients' scores. Interactive software on the website compares the patients' scores to demographically corrected norm scores from the database using advanced multivariate and univariate methods (Huizenga et al., 2007; Huizenga et al., 2016). The ANDI database and accompanying website will simplify normative comparisons, and will provide more sensitive and specific normative comparisons.

In this article, we describe the step-by-step procedure of the ANDI normative database construction, so that the procedure can be replicated in other countries and in other fields of study that also rely on normative comparisons, such as clinical psychology or personnel psychology. We also describe current contents of the ANDI database. Finally, we address the advantages and potential limitations of the ANDI database in comparison to other normative data.

We illustrate these steps using Rey's Auditory Verbal Learning Test (AVLT; Rey, 1958), an internationally well-known test. It is one of the

[1] To avoid confusion: this project is not related to ADNI, which stands for Alzheimer's Disease Neuroimaging Initiative

tests that are also included in the ANDI database. The AVLT measures memory and learning (Lezak et al., 2012; Strauss et al., 2006). In its simplest form participants are presented with a list of 15 nouns, which they are asked to reproduce immediately (in any order). This is repeated five times. Twenty minutes after the five learning trials, there is a delayed recall condition in which participants are asked again which words they remember. Finally, there is a multiple choice recognition condition.

2.3 CONSTRUCTION OF THE ANDI DATABASE

For every neuropsychological test variable included in the ANDI database, a standardized automatized stepwise procedure was followed. A flow chart summarizing all steps can be found in Figure 1. In the following paragraphs, we explain the rationale for the steps and how they were applied.

2.3.1 *Gathering data*

The first step was to collect a large amount of normative data on neuropsychological tests. In cooperation with a group of researchers, the ’ANDI consortium’ (see www.andi.nl for a list of members) was created. The consortium members donated data of healthy control subjects, which they had collected in predominantly clinical research projects. All studies were approved by local ethics committees. All participants had sufficient knowledge of the Dutch language to complete the tests. All data were anonymized and could not be traced back to individual participants.

Example: Data on the (Rey) Auditory Verbal Learning Test (AVLT) from 32 research projects were donated, yielding data from a total of 5121 participants.

2.3.2 *Integrating data*

We created separate files for all neuropsychological tests. Each file contained multiple test variables. Also, the demographic variables age, sex, and level of education, were included for each participant. Only cases with scores on all three demographic variables were included. For each study a unique study identifier was added.

Example: One file for the AVLT was created. In this file data from all test variables were collected. Thus the variable AVLT-1 contained all data on the first trial of the AVLT, the variable AVLT-2 contained data on trial 2, and so on.

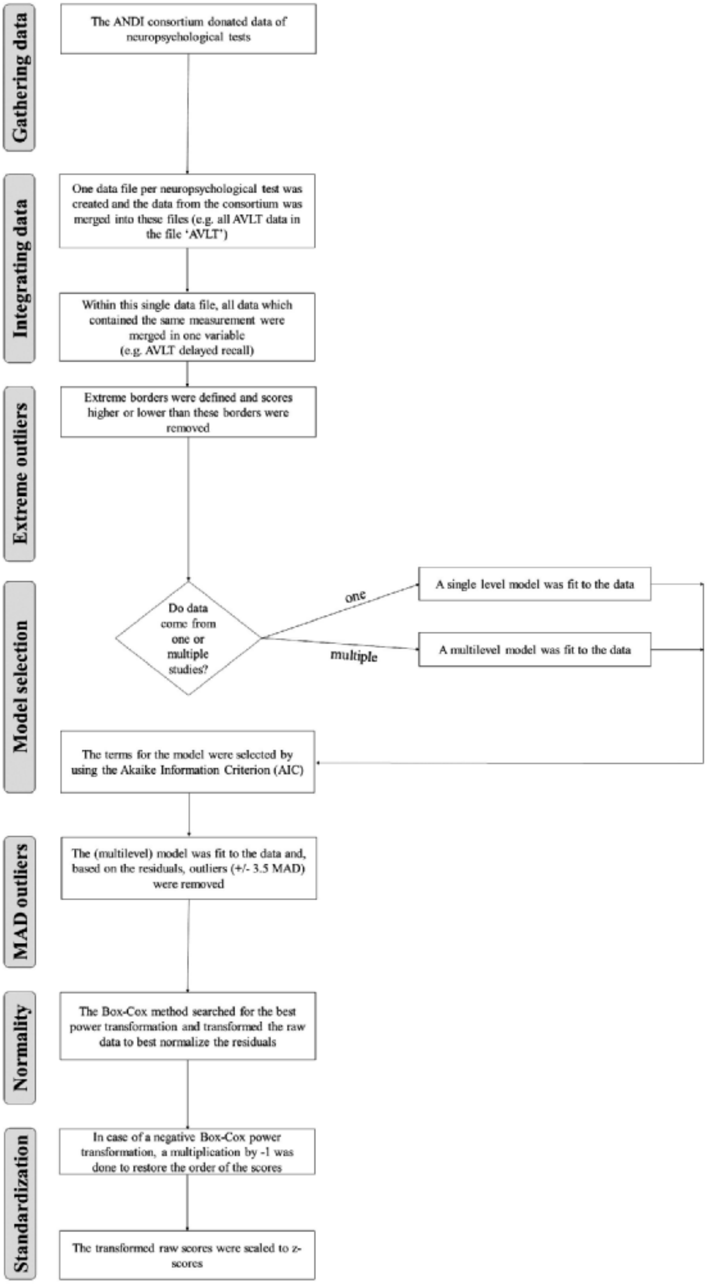
Gathering data
The ANDI consortium donated data of neuropsychological tests
Integrating data
One data file per neuropsychological test was created and the data from the consortium was merged into these files (e.g. all AVLT data in the file 'AVLT')
Within this single data file, all data which contained the same measurement were merged in one variable (e.g. AVLT delayed recall)
Extreme outliers
Extreme borders were defined and scores higher or lower than these borders were removed
Model selection
Do data come from one or multiple studies?
one
A single level model was fit to the data
multiple
A multilevel model was fit to the data
The terms for the model were selected by using the Akaike Information Criterion (AIC)
MAD outliers
The (multilevel) model was fit to the data and, based on the residuals, outliers (+/- 3.5 MAD) were removed
Normality
The Box-Cox method searched for the best power transformation and transformed the raw data to best normalize the residuals
Standardization
In case of a negative Box-Cox power transformation, a multiplication by -1 was done to restore the order of the scores
The transformed raw scores were scaled to z-scores

Figure 2.1: Flow chart describing all steps of the database construction.

2.3.3 *Removing impossible scores*

After merging the data, we checked whether all scores were valid. Invalid scores might be coding errors, or deviant scores observed only in patients with severe pathology. If such invalid scores would not be removed from the database, the variance in scores would be overestimated, which would cause a diminished sensitivity to detect impairments. However, we also wanted the database to be an accurate representation of variability in the healthy population. This implied that the removal criteria should not too strict.

First, we removed the most extreme values. These were scores that were either due to an administrative error or could not come from a healthy participant. For every variable of each neuropsychological test, upper and lower 'extreme borders' were defined. The upper border was set at the maximum possible score. This removed administrative errors. The lower border was set at the worst score a participant can obtain while still deemed cognitively healthy. To this end, we selected the raw score corresponding to the lowest published percentile of the worst performing normative sample. The exact percentile depended on the resolution of the published norm table, but generally a score corresponding to the first percentile was selected. Thus, for a test that has declining scores with increasing age, the raw score that was obtained from the lowest percentile of the oldest participants was defined as the lower extreme border.

If no information from manuals was available, which fortunately was the case for a small number of tests, we asked members of the ANDI consortium to provide acceptable borders. On average 0.48% of scores were removed for the 183 variables. All extreme borders can be found in the ANDI background documentation (www.andi.nl).

Example: The upper border of the AVLT delayed recall is 15. Scores above 15 are impossible and thus were removed. The lower border of AVLT delayed recall was set at 3 after consulting the consortium. Even for the worst performing of the cognitively healthy participants, a score lower than three words was not expected. Such extreme scores could indicate pathology or a typing error, and therefore should be removed. A total of 217 AVLT delayed recall scores (4.5%) fell below the lower extreme border and were removed. No scores exceeded the upper extreme border.

2.3.4 *Model selection*

Next, we used a regression approach to remove demographically corrected outliers. Because a person's neuropsychological test scores depend to some extent on his or her demographic characteristics, not all outlying scores can be found by defining a single criterion value for all scores. For example, scores that are abnormal in young participants may not at all be abnormal in healthy elderly. To define these outliers we, therefore, first wanted to partial out the effects of age, sex, and level of education.

Because the data came from multiple studies, the scores are not strictly independent. For example, some studies may give higher compensation to their participants and these may, therefore, show better scores due to higher motivation. As a second example, some studies may use more stringent exclusion criteria than other studies, and therefore may show higher scores due to the stricter selection of participants. We took variability between studies into account while estimating the effect of age, sex, and level of education using a multilevel regression approach[2] (Curran & Hussong, 2009).

The demographic variables were age in years, sex, and level of education. Level of education was coded on a seven-point scale, which is commonly used in the Netherlands (Verhage, 1964). This scale is similar to UNESCO's ISCED scale (UNESCO, 2012) on which 1 stands for 'no education' and 7 stands for 'university degree'. Although this is an ordinal scale, we treated it as an interval scale and estimated the linear effect of education in order to avoid estimating separate parameters for all levels of education. To determine which effects to include, we first made a selection on the basis of how much demographic information was available, and second, a selection on the basis of which effects were statistically important enough to include in the model. These two selection steps are discussed in more detail below.

part 1: selection of effects based on availability of demographic data.

To estimate the effects of demographic variables, a reasonable range of values on these variables is necessary. However, the range of values was narrow for some variables in the donated data. For example, for some tests only scores from higher educated people were available, which implied that the education effect for these tests could not be estimated.

To find out which effects could plausibly be estimated, we tabulated age, sex and level of education. If the median number of participants in each cell was lower than 5, we considered this too sparse to estimate the corresponding effect. Because age is continuous, we temporarily created age categories, namely individuals younger than 55, aged between 55 and 75 years, and 75+.

Example: In Table 1, an example of this tabulation is given for the AVLT - delayed recall. The effect of sex is estimable, as the minimum cell count across sexes is 2249. The effect of age is considered estimable, as the median cell count across age categories is 1120. Similarly, the effect of education is considered estimable, as the median cell count across education categories is 335.

[2] For variables with data from only one study, a single level regression model was fitted.

Table 2.1: Tabulation of Number of Participants by Sex, Age categories, and Level of Education for the AVLT-Delayed Recall Variable. If the Median (or Minimum in the Case of Sex) Criterion is Not Met for an Effect, this Effect Cannot be Included in the Model.

Sex, N per category	Age, N per category	Level of education, N per category
2249 (Men)	993 (Younger than 55)	17 (1)
2349 (Women)	2485 (55-75 year-olds)	323 (2)
Minimum: 2249	1120 (Older than 75)	119 (3)
	Median: 1120	938 (4)
		1755 (5)
		1111 (6)
		335 (7)
		Median: 335

PART 2: STATISTICAL SELECTION OF EFFECTS TO BE INCLUDED IN THE MODEL. Even if there are sufficient observations to estimate the effect of a demographic variable, it does not necessarily imply that the variable has an effect on the test scores. To determine which effects to include in a regression model, we used a backward selection procedure, removing effects if removal resulted in a lower Akaike Information Criterion (AIC; Cohen, Cohen, West, & Aiken, 2003).

Figure 2 shows the proportions of variables for which effects were included. As can be seen in Figure 2, there were sufficient data to estimate a sex effect for all variables, but in half of the cases, sex was found not to be predictive. Education and age effects were frequently included, if enough data were available to estimate them. The model that was selected for each variable can be found in the ANDI background documentation (www.andi.nl).

Example: For the AVLT-delayed recall the best model included all three effects.

2.3.5 *Removing demographically corrected outliers*

After fitting and selecting the appropriate models to correct for demographic characteristics, we used the residuals rather than the raw scores to decide whether scores were abnormal. These residuals represent the distance of the observed scores from the scores that are expected on the basis of the demographic characteristics. A common criterion for outlying values is three standard deviations from the mean. However, a few outlying scores can increase the standard deviations considerably. Therefore, we used the median absolute devi-

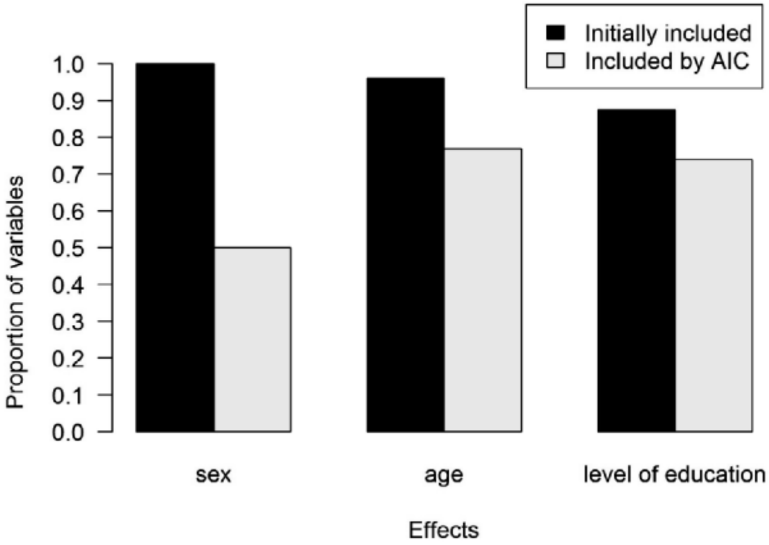

Figure 2.2: Proportion of variables for which the demographic effects were included in the models. In dark gray, effects that could be included after accounting for sample size constraints. In light gray, effects that were included after using the Akaike Information Criterion (AIC) to select effects.

ation from the median (MAD; Leys, Ley, Klein, Bernard, & Licata, 2013), which is more robust to outliers than the standard deviation. As a cutoff criterion, we used 3.5 MAD rather than the more common three standard deviations, as we intended to include as much as possible of the distribution of normal scores. On average 0.53% of scores were removed for the 183 variables.

Example: For the AVLT-delayed recall, no scores exceeded the 3.5 MAD cut off criterion.

NOTE ON THE REMOVAL PROCEDURE. If a participant's score on a test is outlying, one might either remove only this score, remove all of the participant's scores on this test, or remove all of the participant's scores on all tests. We opted for the first possibility, because removing scores on more variables than just the outlying one implies that we can identify the participant's cognitive functioning as the cause of the outlying value, which we cannot. The source may just as well be an administrative error.

2.3.6 *Normality*

The primary aim of the ANDI database is to facilitate normative comparisons. In both univariate and multivariate normative comparison methods, normality of the dependent variables is usually assumed (Crawford & Howell, 1998; Huizenga et al., 2007). Not all neuropsychological test scores, however, are normally distributed. This may be due to effects of demographic variables. For example, if young participants' scores are normally distributed with a low mean reaction time, and if old participants' scores are normally distributed with a high mean reaction time, then the raw scores for both groups combined may be bimodal. However, if the effect of age is partialled out in a regression analysis, and if the residual scores of this regression analysis are used instead of raw scores, such non-normality is no longer an issue. However, residual scores may still be non-normal. For example, some tests show a ceiling effect regardless of the demographic variables. In those cases, a normalizing transformation is recommended to meet the assumption of normality (Crawford, Garthwaite, Azzalini, Howell, & Laws, 2006)}.

Scores are often transformed to normality (Jacqmin-Gadda, Sibillot, Proust, Molina, & Thiébaut, 2007) using transformations such as the square root or the reciprocal. These can both be written as power transformations, raising to the power of 0.5 and -1, respectively. Although these transformations are frequently used, they do not necessarily lead to the best approximation of normality. Therefore, we used the Box-Cox procedure (Box & Cox, 1964; Sakia, 1992) to find the best power transformation. For example, the procedure may find that the best transformation is raising to the power 0.563. The Box-Cox procedure requires a large dataset, which is not often available in neuropsychology (Crawford et al., 2006). Fortunately, the size of the ANDI database allows this Box-Cox procedure.

Because in ANDI, patients will be compared to demographically corrected norms, we wanted the residuals (i.e., scores corrected for the effects of demographic variables) to be normally distributed. The algorithm therefore searches among several power transformations of the raw data (e.g. 0.506, 0.507, 0.508 etc.), and selects the power transformation resulting in the best approximation to normally distributed residuals. The power transformation that was selected for each variable can be found in the ANDI background documentation (www.andi.nl).

The Box-Cox procedure is highly flexible, but our application required a few adjustments. First, all scores have to be larger than 0. Therefore, if there were scores that were either negative or 0, a constant was added (e.g. if the greatest negative value was -5 we added the constant 5.001) to make all scores positive. Second, if the best power transformation turned out to be negative, raising the raw scores to this power flipped the order of values, i.e. the worst scores became the best and vice versa. To reverse this change of ordering, these transformed values were multiplied by -1 to restore their original order. Third, we included study as a predictor in the regression model, because we wanted the residuals to be normal within every study instead of normal over studies. Fourth, power transformations may result in tiny or huge values, which may be difficult to interpret. Therefore, we first Box-Cox transformed all scores, and then standardized all these transformed scores to the familiar z-scale with mean 0 and standard deviation 1. Finally, all standardized transformed z-scores were merged into a single dataset to create the final ANDI database.

Example: For AVLT-delayed recall, the best Box-Cox power transformation was 0.75, implying that when raw scores on AVLT-delayed recall were raised by the power 0.75, the residuals were as normally distributed as possible. In Figure 3 and 4, it can be seen that the residuals were somewhat skewed before transformation and were neatly normally distributed after transformation.

When a patient's scores are compared to the scores in the database, the patient's scores are automatically transformed by the ANDI website's software using the same procedure.

2.3.7 *Model evaluation*

FIT TO DATA After outlier removal, transformation, and standardization, the (multilevel) regression models were fitted again. This was done to get parameter estimates on the new standardized transformed scale. To evaluate whether the model fitted the raw data well, predictions from the model had to be destandardized and transformed back to the original scale. These back-transformed model predictions were plotted together with the raw data for visual inspection of model fit.

Example: In Figure 5, the raw scores on the AVLT delayed recall variable are plotted as a function of age, sex, and level of education. All raw scores lie between 3 and 15, as extreme outliers have been removed. There are many data points for education levels 2 through 7, but relatively few for education level 1 . All effects were included in the model. This can be observed in Figure 5. The effect of age indicates that scores decrease as participants get older. It can also be observed that men do slightly worse than women, and that scores increase with level of education.

In Figure 6, between and within study variance is plotted for the variables originating from multiple studies. It can be seen that between study variance exists for most of the variables, although between study variance was generally lower than within study variance.

2.4 CONTENTS OF ANDI

ANDI currently contains data of 26,635 healthy participants on 43 neuropsychological tests from different cognitive domains. As an ex-

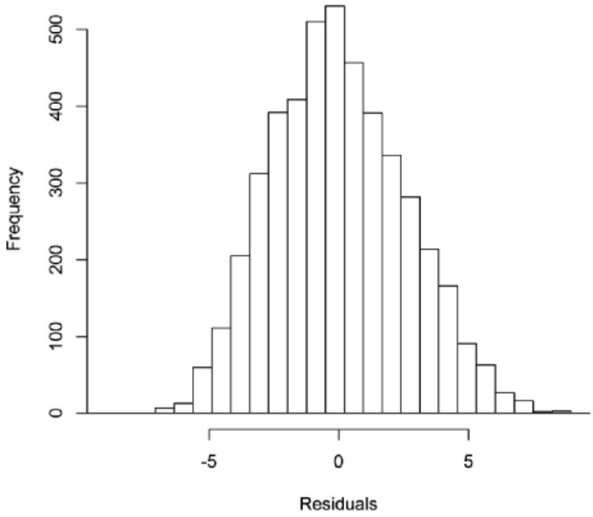

Figure 2.3: Distribution of the residuals of the model fitted to the AVLT delayed recall variable before power transformation.

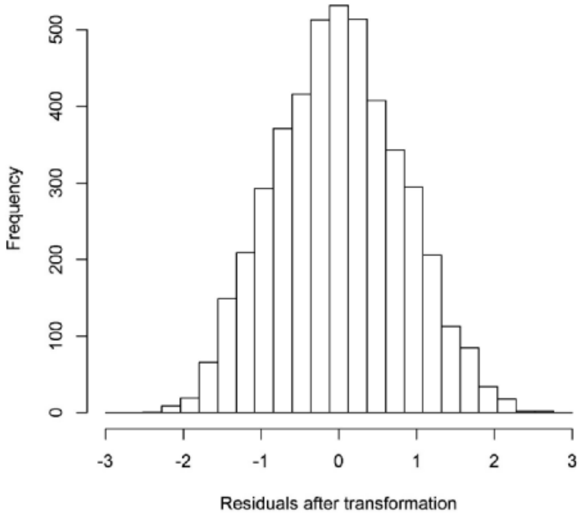

Figure 2.4: Distribution of the residuals of the model fitted to the AVLT delayed recall variable after the power transformation of 0.75, and after standardization

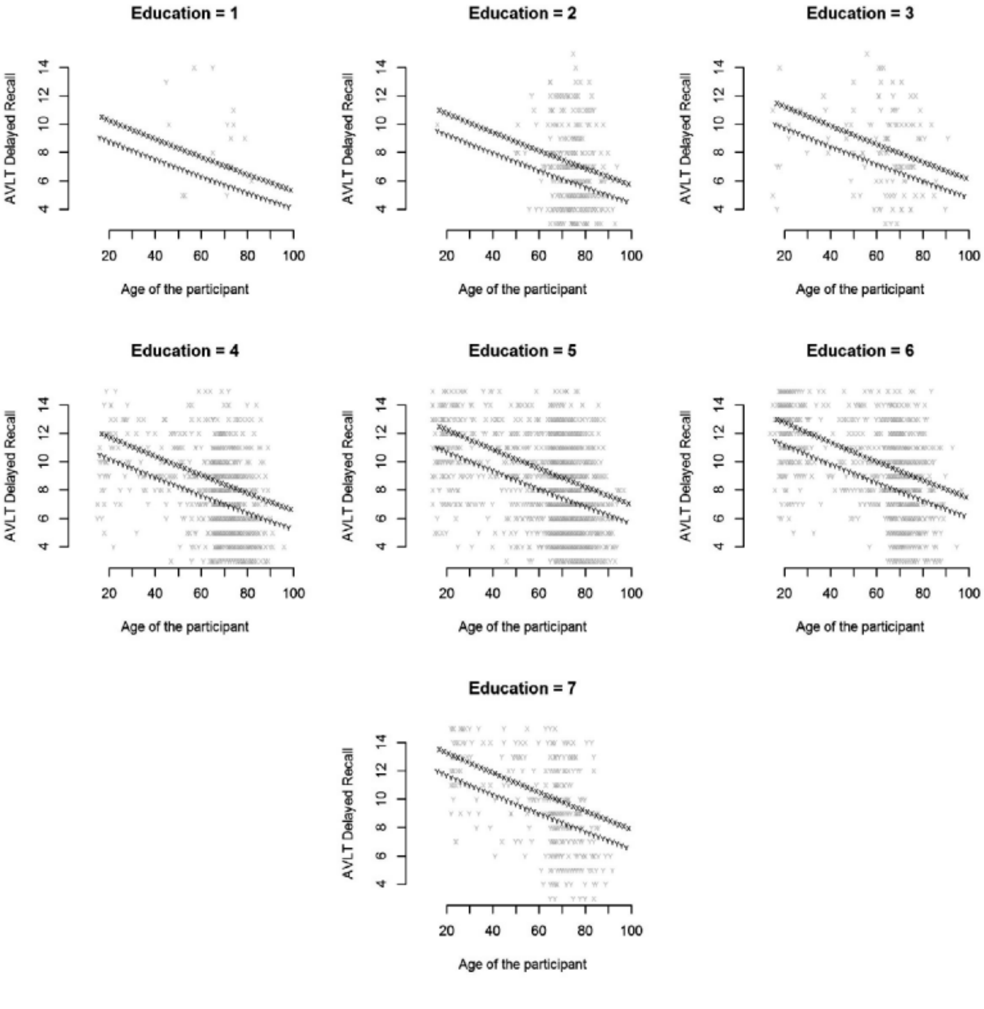

Figure 2.5: Raw scores on the AVLT delayed recall variable are plotted against age. Separate plots were made for the different levels of education. Men are depicted with the letter y and women with x.

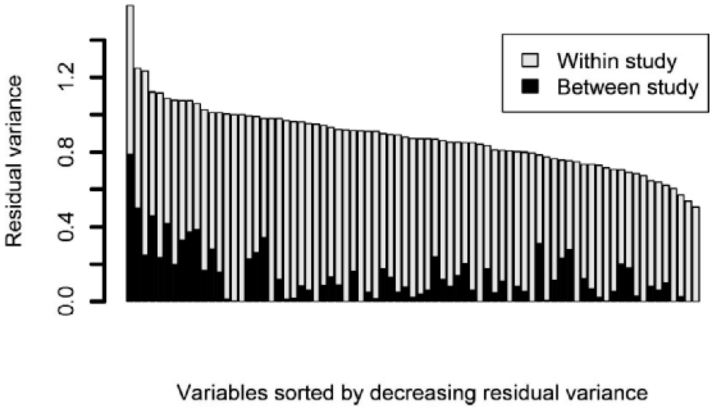

Figure 2.6: Partitioning of total residual variance for variables that were administered in multiple studies. Dark gray portions of the bars are the residual variance due to between study differences. Light gray portions of the bars are the residual variance due to within study/between participant differences.

ample, Table 2 lists a selection of variables currently included in the database (the complete list is available on www.andi.nl).

2.5 DISCUSSION

We described the steps to prepare the ANDI database for normative comparisons in neuropsychology. First, data were gathered from the ANDI consortium. Second, data from neuropsychological tests were merged. Third, we removed scores that could not come from cognitively healthy participants using extreme borders. Fourth, to determine for which demographic effects to correct, we selected only effects for which we had enough data and only included the effect if this was necessary according to the AIC. Fifth, after a model had been defined, we removed scores that were outlying after correction for demographic characteristics. We did this by removing scores that differed more than 3.5 MAD from the median. Sixth, because normative comparison procedures assume normality of score distributions, we used the Box-Cox procedure to search for a power transformation that, when applied to the raw data, optimally normalized the residuals after the demographic correction. These steps were applied for every variable of every neuropsychological test included in the database.

Table 2.2: Example Variables per Neuropsychological Test.

Example variable	N studies	N in ANDI	Age range	% Men	Education range
Executive functions					
Letter Fluency (3 letters)	23	2897	17-97	48	1-7
Semantic Fluency (animals)	27	5783	17-96	40	1-7
BADS (Zoo map total)	6	398	17-86	43	1-7
Attention and Working Memory					
Trail Making Test A	37	3320	8-97	47	1-7
Trail Making Test B	37	3254	8-97	47	1-7
Stroop (Word in seconds)	30	2147	16-91	43	1-7
Stroop (CW Interference in seconds)	30	2132	16-91	43	1-7
Visuospatial					
Judgment of Line Orientation (raw score)	1	69	40-80	54	3-7
Memory					
RAVLT (delayed recall)	29	4598	14-97	49	1-7
RBMT (prose 1 delayed recall)	8	396	17-89	44	1-7
RCFT (delayed recall)	5	502	17-86	48	1-7
WAIS III Coding	9	1734	15-92	49	1-7
Language					
Boston Naming Test (long version)	5	400	17-89	40	1-7
Intelligence					
Dutch Adult Reading Test (raw score)	26	2171	16-96	42	1-7
Raven CPM (A+B)	2	4020	55-94	48	1-7

2.5.1 *Benefits of the ANDI database*

The ANDI database and infrastructure offer a number of advantages over existing normative data published in test manuals.

MORE APPROPRIATE NORMS First, the ANDI normative data have been gathered roughly in the past 20 years which make them more applicable than data that have been gathered longer ago. Because the database is internet-based, and because the ANDI construction procedure is highly automatized, it will be possible to keep the norms up-to-date by regularly adding new data and rerunning the ANDI construction procedure. Second, the ANDI database contains a considerable amount of data for old (80+) participants, making normative comparisons for this group also feasible. Third, because the data have been donated by universities and hospitals in the Netherlands and Flemish Belgium, all norms come from a population similar to patients in these countries. Fourth, scores in ANDI are corrected for the effects of age, sex, and level of education. This is an improvement over many published normative data which are typically corrected for age only. Fifth, in many traditional norms, age is not treated as a continuous variable, but is divided into arbitrary age categories. This implies that when one shifts from one age category to the next, the interpretation of the test score may change abruptly. Because in our regression approach age is treated as a continuous variable, such leaps between groups do not occur (Testa et al., 2009). Sixth, for many test variables, the ANDI norms are based on large numbers of participants (e.g., thousands) making them more precise than many existing neuropsychological norms.

NORMATIVE COMPARISONS WITH MULTIVARIATE DATA Another unique aspect of ANDI as a normative database is that many participants in the database have completed multiple tests. This allows multivariate normative comparisons, which have increased sensitivity to detect cognitive impairment (Huizenga et al., 2007). Multivariate norms are currently often lacking so that multivariate normative comparisons cannot be broadly applied in clinical practice. Likewise, multiple testing corrections for univariate normative comparisons which also require multivariate normative data (Huizenga et al., 2016), and normative comparisons that compare differences between test scores within one patient (Crawford & Garthwaite, 2002), become feasible. With the ANDI database and the accompanying website, such comparisons can be routinely applied.

EXPORTABLE INFRASTRUCTURE The software of the ANDI infrastructure will be freely available for researchers to be applied to other data sets. If researchers collect their own control datasets, the highly automatized procedure for merging, standardization and correction of the scores described here could be carried out (all code is provided on https://github.com/JAvRZ/andi-dataprocessing). In this way, versions in other countries and other fields of study (such as clinical psychology or medicine) can be set up.

2.5.2 *Potential limitations of the ANDI database*

It is important to mention potential limitations of the ANDI database. First, ideally a normative database is based on a random sample. Although some included studies indeed sampled randomly from the population, others used convenience samples, e.g. they used family members of patients as controls. However, note that the effects of age, sex and level of education were included in the models, thereby removing potential confounding effects of convenience sampling. Second, the sample should ideally be from a cognitively healthy population. Indeed, some donated studies assured that pathology was absent in the control sample, however others used more lenient inclusion criteria. We tried to mediate this problem by excluding impossible and outlying scores.

2.5.3 *Concluding remark*

Although our primary goal is to make a contribution to neuropsychological assessment, we also strive for broader applications. The highly automatized ANDI construction procedure software is freely available, allowing others to build their own diagnostic infrastructure. Creating such database-supported infrastructures can be an important innovation in healthcare and health research as it may provide better norms and more advanced diagnostic procedures. In research projects, it may replace collecting data from control subjects if the control data can be obtained from the database. This shows once more that data sharing has great potential. Newly created databases –like ANDI– provide valuable new resources while not putting any additional burden on healthy controls.

3

MULTIVARIATE NORMATIVE COMPARISONS USING AN AGGREGATED DATABASE

3.1 ABSTRACT

In multivariate normative comparisons, a patient's profile of test scores is compared to those in a normative sample. Recently, it has been shown that these multivariate normative comparisons enhance the sensitivity of neuropsychological assessment. However, multivariate normative comparisons require multivariate normative data, which are often unavailable. In this paper, we show how a multivariate normative database can be constructed by combining healthy control group data from published neuropsychological studies. We show that three issues should be addressed to construct a multivariate normative database. First, the database may have a multilevel structure, with participants nested within studies. Second, not all tests are administered in every study, so many data may be missing. Third, a patient should be compared to controls of similar age, sex, and educational background rather than to the entire normative sample. To address these issues, we propose a multilevel approach for multivariate normative comparisons that accounts for missing data and includes covariates for age, sex, and educational background. Simulations show that this approach controls the number of false positives and has high sensitivity to detect genuine deviations from the norm. An empirical example is provided. Implications for other domains than neuropsychology are also discussed. To facilitate broader adoption of these methods, we provide code implementing the entire analysis in the open source software package *R*.

3.2 INTRODUCTION

In neuropsychological assessments, a battery of tests is administered to a patient to determine whether his or her cognitive functions are impaired (Lezak et al., 2012; Strauss et al., 2006). Tests within these batteries are designed to assess the patient's memory, attention, language capacities or other functions. To interpret the patient's scores, these scores have to be compared to the distribution of test scores in healthy controls. Such a comparison is called a normative comparison. A clinical neuropsychologist may use one standard deviation below

[0] Published as: Agelink van Rentergem, J. A., Murre, J. M. J., & Huizenga, H. M. (2017). Multivariate normative comparisons using an aggregated database. *PLoS ONE*, *12*, 1-18.

the mean as a criterion for impairment (Brooks, Iverson, & White, 2009). When a patient's test scores are found to be below normal, this helps the neuropsychologist characterize the patient's cognitive deficit, and may guide differential diagnosis and treatment.

In neuropsychological research, normative comparisons can be used in a similar way. For example, if a patient and a control group are studied, normative comparisons can be made for each patient in the patient group, with the distribution of test scores in the control group as the reference. In this manner, new variables can be constructed that index whether patients deviate from the norm or not. Such indices may for example be used to assess whether a new treatment, as compared to a waiting list condition, reduces the number of patients who deviate from the norm (Kraemer et al., 2003).

Normative comparisons are generally conducted for each test separately: The patient's test score is compared to the distribution of test scores for that specific test. This is the univariate approach to normative comparisons. An alternative approach is to compare the patient's profile of test scores to the multivariate distribution of test scores. This is the multivariate approach to normative comparisons (Huba, 1985; Crawford & Allan, 1994; Huizenga et al., 2007; Grasman, Huizenga, & Geurts, 2010). Multivariate comparisons have been shown to be more sensitive than univariate comparisons to detect deviations (Su et al., 2015). For example, profiles of high scores on some tests and low scores on other tests, or profiles with many scores that are only a little below normal, are readily detected (Huizenga et al., 2007). An additional advantage is that no correction for multiple comparisons is required (Huizenga et al., 2016), because only a single multivariate comparison is conducted. Multivariate normative comparisons have been applied in the study of disorders as diverse as Parkinson's disease (Smeding et al., 2010; Castelli et al., 2010, Broeders et al., 2013), stroke (Phaf, Horsman, van der Moolen, Roos, & Schmand, 2010), prosopagnosia (Valdés-Sosa et al., 2011), bacterial meningitis (Schmand et al., 2010) and HIV-associated neurocognitive disorder (Cohen et al., 2014; Su et al., 2015).

Multivariate normative comparisons for two hypothetical situations are illustrated in Figure 1.. In the left panel, the correlation between the memory test score and language test score is 0. In the right panel, the correlation is 0.7. Univariately, the test scores of a hypothetical patient do not deviate, in both panels. Multivariately, the combination of the above average score on the language test, and the below average score on the memory test does not deviate in the left panel, but does deviate in the right panel. In other words, in the right panel, the multivariate comparison shows that the memory score is indeed weak, given the strength of the language score. An experienced clinician may recognize this deviating profile given his/her intuition on the correlation between test scores in the norm group. He/she may

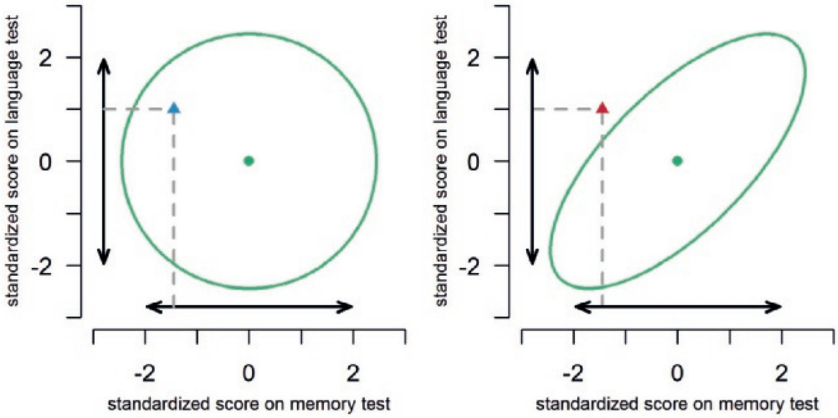

Figure 3.1: Illustration of multivariate normative comparisons in a situation with scores on two neuropsychological tests. The double-headed arrows denote the 95% univariate intervals. The ellipses denote the 95% multivariate region. The dots denote the mean score in the norm group. The triangles depict a patient's scores. In the left panel, tests are uncorrelated (r = 0.0). In the right panel, tests are correlated (r = 0.7).

be able to decide without using a formal multivariate procedure that the low score on one test together with the high score on the other test is a cause for concern. However, in situations with more than two tests, or situations that are less familiar to the clinician, such decisions will become more difficult. A formal multivariate comparison should then fare better than an informal one, and is likely to promote more accurate diagnostic decisions.

An important drawback of this multivariate method is that multivariate normative data are required, because it is necessary to estimate the covariance of test scores within the norm group (Grasman et al., 2010; Huizenga et al., 2007). As test developers typically focus on one test, or at most a few tests at a time, these multivariate normative data are not often available. A solution might be to obtain normative data from a neuropsychological study in which a clinical sample has been compared to a healthy control sample on multiple neuropsychological tests. However, a single neuropsychological study will not provide normative data on all neuropsychological tests as, in any single study, only a limited number of tests are administered. Fortunately, by combining data from multiple neuropsychological studies, a dataset can be established that provides all required information. This is the approach that was chosen in a recently started project (www.andi.nl). In this project, a composite normative dataset has been constructed from healthy control data provided by several research institutes. In the following we outline the issues that arise in the construction of such a database.

Table 3.1: Example of a Missing Data Pattern, Where 1 = Available, and 0 = Missing. For Each Test, and Each Study, There Are Scores Missing, Although All Tests Co-occur at Least Once.

	Test 1	Test 2	Test 3
Study 1	1	1	0
Study 2	1	0	1
Study 3	0	1	1

First, test scores may differ from study to study. Although neuropsychological tests are highly standardized, subtle differences between studies may arise due to the design of the studies. Such differences might for example be caused by differences in incentives that are given to participants, or by differences in the order of test administration. Second, certain tests are administered in one study but not in others (cf. Table 1). That is, for many participants, data will be missing on those tests that were not administered in the study they participated in. The common approach of listwise deletion discards all participants with incomplete data (Schafer & Graham, 2002), and would result in no participants at all.

These two issues, missing data and differences between studies, can adequately be handled by multilevel modeling. Multilevel modeling can account for variance between studies (Tabachnick & Fidell, 2007) and multilevel modeling allows for missing values (Schafer & Yucel, 2002). Therefore, the present paper provides a multilevel modeling extension of the multivariate approach to normative comparisons.

In making normative comparisons, it is important to correct for background variables that might influence scores. For instance, age may affect reaction times in such a way that a reaction time that implies brain damage in young adults may not be particularly uncommon in a very senior but healthy population. Similarly, a score that implies mild cognitive impairment in highly educated individuals may not be uncommon in healthy individuals with a lower education. Sex usually is less influential, but can make a difference in certain verbal tests, on which women do slightly better than men, and in some visuospatial tests, on which women may do slightly worse (Lezak et al., 2012). Because of the importance of these background variables, test manuals often contain extensive norm tables to which the score of a patient can be compared. For every background variable that is added as a potential predictor, a new dimension is added to the table.

As an alternative to norm tables that are split for different background variables, regression-based norms are becoming increasingly common (Crawford & Howell, 1998; Crawford et al., 2006). Instead of defining subgroups, participants are compared to the predicted score of a regression equation, in which test scores are regressed on

Table 3.2: Simulated Example of a Multilevel Dataset with One Row per Test Score. study Indicates Study Number; ID Indicates Participant Number; age, sex, and education are Background Variables; z(1), z(2) and z(3) are Indicator Variables; test Indicates Test Number and score Indicates the Score on the Test with that Number.

study	ID	age	sex	education	z(1)	z(2)	z(3)	test	score
1	1	-2.21	-1	3.68	1	0	0	1	0.08
1	1	-2.21	-1	3.68	0	1	0	2	1.59
1	2	22.79	1	-0.32	1	0	0	1	0.72
1	2	22.79	1	-0.32	0	1	0	2	2.06
2	1	-25.21	1	0.68	0	1	0	2	0.19
2	1	-25.21	1	0.68	0	0	1	3	1.26
2	2	-11.21	1	1.68	0	1	0	2	0.04
2	2	-11.21	1	1.68	0	0	1	3	-0.29
2	3	3.79	-1	0.68	0	1	0	2	-0.65
2	3	3.79	-1	0.68	0	0	1	3	-0.51

background variables such as age, sex and educational background (Testa et al., 2009; Parmenter, Testa, Schretlen, Weinstock-Guttman, & Benedict, 2010). In order to correct for these background variables in a regression-based manner, we add the background variables to the multilevel procedure as well.

In this paper, we first describe the multilevel approach to multivariate normative comparisons. We then use Monte Carlo simulations to test the efficacy of this approach in terms of false positives and in terms of sensitivity to genuine deviations from the norm. We demonstrate the application of the method. We conclude by discussing assumptions and by suggesting some future directions.

3.3 METHODS

A multilevel analysis requires that the data are structured such that every row of the dataset represents a single test score for one participant. An example with simulated data for three tests is given in Table 2.

In Table 3, the model specification is given. The model consists of three levels: the level of test scores (abbreviated to tests, although some tests may produce multiple scores), the level of participants and the level of studies.

At level 1, scores are expressed as a function of so-called indicator variables. These variables indicate to which test the dependent variable refers. If the indicator variable z(1) is 1, the dependent variable test score refers to Test 1, if z(2) is 1, the variable test score refers to

Table 3.3: Model Specification for a Multilevel Model with Three Tests, and Three Background Variables (age, sex, and level of education), Including Specification of Between and Within Study Covariance Structures.

Level 1 (test : i)

$$y_{ijk} = \beta_{1jk}z(1)_{ijk} + \beta_{2jk}z(2)_{ijk} + \beta_{3jk}z(3)_{ijk}$$

Level 2 (person : j)

$$\begin{aligned} \beta_{1jk} &= \phi_{10k} + \phi_{11k}age_{jk} + \phi_{12k}sex_{jk} + \phi_{13k}education_{jk} + \epsilon_{1jk} \\ \beta_{2jk} &= \phi_{20k} + \phi_{21k}age_{jk} + \phi_{22k}sex_{jk} + \phi_{23k}education_{jk} + \epsilon_{2jk} \\ \beta_{3jk} &= \phi_{30k} + \phi_{31k}age_{jk} + \phi_{32k}sex_{jk} + \phi_{33k}education_{jk} + \epsilon_{3jk} \end{aligned}$$

Level 3 (study : k)	Intercept	Age	Sex	Education
	$\phi_{10k} = \gamma_{100} + \nu_{10k}$	$\phi_{11k} = \gamma_{110}$	$\phi_{12k} = \gamma_{120}$	$\phi_{13k} = \gamma_{130}$
	$\phi_{20k} = \gamma_{200} + \nu_{20k}$	$\phi_{21k} = \gamma_{210}$	$\phi_{22k} = \gamma_{220}$	$\phi_{23k} = \gamma_{230}$
	$\phi_{30k} = \gamma_{300} + \nu_{30k}$	$\phi_{31k} = \gamma_{310}$	$\phi_{32k} = \gamma_{320}$	$\phi_{33k} = \gamma_{330}$

Combined (substitution of level 3 into 2, and level 2 into 1)

$$\begin{aligned} y_{ijk} = \quad & (\gamma_{100} + \gamma_{110}age_{jk} + \gamma_{120}sex_{jk} + \gamma_{130}education_{jk} + \nu_{10k} + \epsilon_{1jk})z(1)_{ijk} + \\ & (\gamma_{200} + \gamma_{210}age_{jk} + \gamma_{220}sex_{jk} + \gamma_{230}education_{jk} + \nu_{20k} + \epsilon_{2jk})z(2)_{ijk} + \\ & (\gamma_{300} + \gamma_{310}age_{jk} + \gamma_{320}sex_{jk} + \gamma_{330}education_{jk} + \nu_{30k} + \epsilon_{3jk})z(3)_{ijk} \end{aligned}$$

Covariance Matrix Within				Covariance Matrix Between			
	Test A	Test B	Test C		Test A	Test B	Test C
Test A	$var_{\epsilon_{1jk}}$			Test A	$var_{\nu_{10k}}$		
Test B	$cov_{\epsilon_{2jk},\epsilon_{1jk}}$	$var_{e_{2jk}}$		Test B	0	$var_{\nu_{20k}}$	
Test C	$cov_{\epsilon_{3jk},\epsilon_{1jk}}$	$cov_{\epsilon_{3jk},\epsilon_{2jk}}$	$var_{\epsilon_{3jk}}$	Test C	0	0	$var_{\nu_{30k}}$

test 2, etc. A similar method using indicator variables for multivariate data analysis has been described before (Goldstein, 1995; Bauer, Preacher, & Gil, 2006).

At level 2, the effects of a participant's background variables, that is, age, sex, and educational background, are introduced. The level 2 model also includes the error terms ϵ_{ijk}, which denote deviations of an individual's observed test scores to that predicted by the model for that particular study.

At level 3, differences between studies are introduced by adding error terms v to the intercept of each test. Note that the effects of age, sex, and educational background are constrained to be the same in different studies, as it is unlikely that these effects differ between studies. This constraint can however easily be relaxed by adding error terms to those effects as well.

Substituting level 3 into level 2, and level 2 into level 1 yields the combined model (cf. Table 3). In this model, γ_{100} denotes the intercept of the first test. The interpretation of intercepts is dependent on the scaling of background variables. If age and education are centered on their mean and sex is contrast coded, the intercept γ_{100} refers to the scores on the first test for an "average" participant: of average age, not of a specific sex, with an average educational background. The parameters γ_{110}, γ_{120} and γ_{130} denote the effects of age, sex, and educational background on the first test. In addition to these so-called fixed effects, the model also yields estimators of random effects: the covariance matrix of within study errors ϵ and the covariance matrix of between study errors v.

No constraints were imposed on the covariance structure of within study errors ϵ (cf. Table 3). Modeling each of the covariances between variables separately can account for both dependencies between variables within tests, and between variables that belong to different tests. Also, measurements can freely covary both positively and negatively. The covariance matrix of within study errors was constrained to be equal over studies, as is common in multilevel modeling (Tabachnick & Fidell, 2007).

As it is unlikely that test scores of "average" participants covary at the between study level, we imposed the constraint that between study errors v did not covary (cf. Table 3). This constraint could be relaxed by adding these covariances to the model as well.

As mentioned in the introduction, one of the advantages of multilevel modeling is the handling of missing values. More specifically, multilevel models do not require that every participant has completed an equal number of tests. Multilevel models can be estimated with Full Information Maximum Likelihood (FIML) which uses all available information from each case (Dempster, Laird, & Rubin, 1977; Enders & Bandalos, 2001). For FIML to result in correct parameter estimates, the missing data mechanism should be ignorable, i.e., the fact that an observation is missing should not be due to the value of that particular observation (Schafer & Graham, 2002). In the present case, missing data is due to the study design (Graham, Taylor, Olchowski, & Cumsille, 2006), and not due to the values of test scores that participants achieve. Since the participants that are pooled are all healthy, and the tests can be completed easily by healthy participants, missing data within studies should not occur systematically. Therefore, the missing data mechanism can be classified as ignorable, and FIML will yield adequate estimates.

In sum, multilevel modeling can be used to combine the results of multiple studies, even if data are missing, and it can incorporate background variables. Next, we indicate how multilevel models can be combined with multivariate normative comparisons to analyze whether an individual deviates from a composite normative database.

The multivariate normative comparison uses a version of Hotelling's T^2 statistic that is adapted for normative comparisons. If there are no background variables, the equation for this T^2_{norm} is (Huizenga et al., 2007; Grasman et al., 2010):

$$T^2_{norm} = \frac{1}{(n+1)/n} \frac{n-p}{(n-1)p} (\bar{y} - x)' C^{-1} (\bar{y} - x) \qquad (3.1)$$

where n is the number of participants in the norm group, p is the number of tests, $\bar{y}$ is a vector of length p containing the mean scores for every test in the norm group, x is a vector of length p containing the patient scores for every test, prime $'$ denotes transposition, C is the p by p covariance matrix of the test scores in the norm group, and C^{-1} is the inverse of this covariance matrix.

Looking up T^2_{norm} in the F-distribution with p numerator degrees of freedom, and $n - p$ denominator degrees of freedom, yields a p-value corresponding to the probability that the patient would obtain this profile of scores (or a more extreme one) if he belongs to the same population as the norm group (Grasman et al., 2010; Huizenga et al., 2007). If this probability is very small, for example smaller than 0.05, the patient's profile of scores is said to be deviating.

This normative comparison is two-sided, as both overall positive and overall negative deviations are considered abnormal. A one-sided variant has also been developed (Follmann, 1996; Huizenga et al., 2007). In one-sided testing, all tests have to be standardized to bring them on the same scale. It is then decided that an individual is deviating from the norm if two conditions are satisfied: (1) the sum of deviations over tests is in the expected direction, and (2) the p-value does not exceed 0.10.

To account for the multilevel structure in the normative database, we make three adjustments to the multivariate normative comparisons method. First, the covariance matrix C is now the sum of two covariance matrices: the within study covariance matrix and the between study covariance matrix. Second, y now denotes the normative scores predicted given an age, sex, and level of education that matches that of the patient. Third, the degrees of freedom have to be adjusted, as, in case of missing data, participants do not contribute information to the estimation of all parameters, and as individuals are nested within studies and thus observations are not completely independent (Tabachnick & Fidell, 2007).

There is no consensus on how degrees of freedom should be computed and different software packages use different methods (Bolker et al., 2009). We use the method implemented in the multilevel modeling software package *nlme* (Pinheiro & Bates, 2000), which for our case equals the number of observations - (number of studies + number of estimated effects + 1).

Similar to the issue of determination of degrees of freedom, determination of the n to be used in equation (1) is not straightforward when dealing with nested and missing data. Fortunately, once n becomes moderately large (above 100), even large differences in choice of n are of little influence. We set n equal to the total number of participants.

3.4 simulations

In simulation study 1, we investigated the effect of ignoring between study variance on the false positive rate. We did this by fitting models both with and without between study variance. In simulation study 2, we investigated the effect of missing data on false positive rate and sensitivity. In simulation 2, scores on certain tests were deleted for all participants in a study, as if the researchers in that study had decided not to administer that test.

3.4.1 *Methods*

The settings for the simulation studies are given in Table 4. In simulation study 2, either 0%, 40% or 70% of the data was made missing. Missing data was introduced by deleting data according to the pattern in Figure 2. The 0% condition is intended not as a control condition, but as a check of multilevel normative comparisons, without the added complication of missing data. Because of the nature of the aggregate database, 0% missing data will never be encountered in real settings. If only regularly administered tests were included in the database, only 40% to 70% missing data should be achievable. However, if all possible neuropsychological tests were included, the percentage missing test scores should be much higher; this would not allow the current model specification and normative comparison methods. This limitation is discussed further in the discussion section. Ten tests were used in the simulation: Twelve tests is the average

Table 3.4: Settings Used in the Two Simulation Studies.

Settings	
Number of tests	10
Number of participants per study	50
Number of studies	30
Percentage of test scores missing	Simulation 1: 0 %
	Simulation 2: 0 %, 40% or 70%
Number of simulations	1000 per condition

Table 3.5: Parameter Values in the Two Simulation Studies.

Parameters	
Intercepts	20
Age effect	-0.125
Sex effect	0.5
Education effect	1.25
Residual variance of test scores within studies	25
Residual correlation between test scores within studies	0.4
Residual variance of test scores between studies	5
Residual correlation between test scores between studies	0.0

number of tests that a neuropsychologist uses (Rabin, Paolillo, & Barr, 2016).

The parameter values for the two simulation studies are given in Table 5. The ANDI database was used to set the sample sizes of studies and the number of studies. The ANDI database was also used to estimate the effect sex, age and level of education would have on test scores. The simulation settings (see doi.org/10.5281/zenodo.321858) were based on these estimates. Information on the ANDI database (which groups contributed, how many studies and participants are available per test variable etc.) is presented in the documentation on www.andi.nl. Another large Dutch sample was examined to verify that effects as observed in the ANDI database can be considered representative (Murre, Janssen, Rouw, & Meeter, 2013). The effects of background variables were all assumed to be linear. A parameter of -0.125 for age indicates for example that for every year that a participant increases in age, the participant on average achieves a score that is 0.125 points lower. The variance between studies was assumed to be small compared to the variance between participants within studies.

In both simulation studies, patient data were simulated with the same parameters that were used to simulate normative data, on the

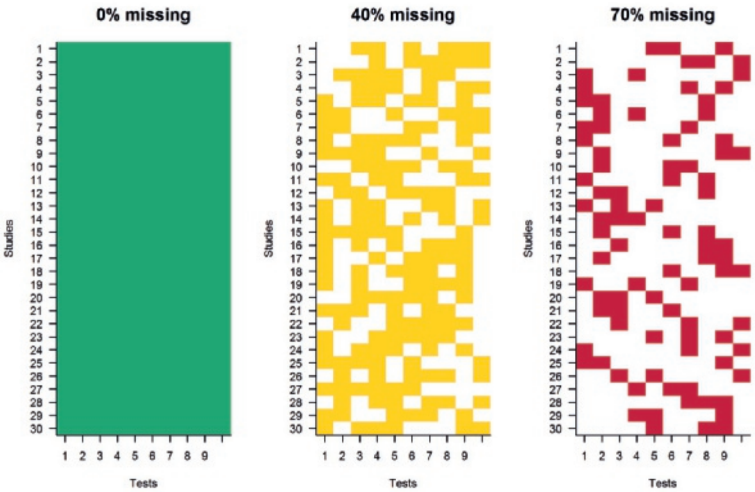

Figure 3.2: Missing data patterns for the 0%, 40% and 70% missing data conditions, with studies on the y-axis and tests on the x-axis. Colored boxes are non-missing test scores, white boxes are missing test scores.

understanding that patients' scores differed from the scores in the norm group on 0, 1, 2, 5 or 9 tests. These deviations were introduced by subtracting two standard deviations (computed from the total variance) of the test scores in the norm group from the patient's simulated test scores. So if patients truly deviated, they did so in a negative way. Two standard deviations could be considered the difference between patients and the norm group that is maximally interesting from a statistical perspective: Patients with much more extreme scores are easily recognized as being deviating, and patients with much less extreme scores are probably non-deviating. A 2 standard deviation difference is however a large difference in neuropsychological terms. Research has shown that 1 and 1.5 standard deviations are common for impairments that are secondary to a particular disorder, for example for attention problems that accompany major depression (Zakzanis, Leach, Kaplan, 1998).

In applying the multivariate comparison, false positive rate was defined as the fraction of simulations in which a significant multivariate difference was observed in conditions in which there were no simulated differences. Sensitivity was defined as the fraction of simulations in which a multivariate difference was observed, in conditions in which simulated differences were present.

The multivariate results were contrasted with results of univariate comparisons. For the univariate comparisons, we recorded whether *any* of the patients' scores, using an alpha of 0.05, deviated significantly from the norm univariately. This implies that in the power condition, we did not require that the deviation corresponded to any of the simulated deviations. This definition keeps results comparable between univariate and multivariate results, but works in favor of univariate comparisons: They do not need to be correct to be sensitive.

In the case of no simulated deviations, the rate of finding at least one deviation is known as the familywise error rate. It has been shown that the familywise error rate becomes much too high if multiple comparisons are made (Huizenga et al., 2007). Therefore, corrections can be applied, such as the Bonferroni correction, which divides the criterion for significance by the number of comparisons. Therefore, we compared the results that were obtained using the multivariate comparisons to univariate comparisons that were either uncorrected or Bonferroni corrected.

All comparisons were one-sided, as clinicians are generally only interested in patients' performance being worse than in the norm group. This means that we used a p-value of 0.10 for the multivariate comparison as our criterion value, with the added criterion that the summed difference is in the expected direction, as described in the method section. Given these two criteria, we expect the overall proportion of significant deviations to equal 0.05 if no differences were simulated. For the univariate comparison, we used 0.05 as our one-sided criterion.

A critical p-value of 0.05 or equivalently a 95% confidence interval is often used in scientific research, but not in clinical practice. In clinical practice, more lenient criteria, such as 1 SD or 1.5 SD below the mean are common. In fact, research has shown that sensitivity and specificity may be optimal with such a 1.5 SD criterion (Dalrymple-Alford et al., 2011). However, in applications of the multivariate normative comparison, the 95% confidence interval has been shown to be sensitive to deviations, even in comparison to univariate results with more lenient criteria (Su et al., 2015). Therefore, the 0.05 criterion was used in these simulations as well.

We fitted the multilevel models using the software package *nlme* (Pinheiro & Bates, 2000), because it is flexible in specifying covariance structures both for the ϵ and v terms. *R* code that can be used to perform the entire analysis including the multivariate normative comparison can be found in the supporting information in the online version.

3.5 results

3.5.1 *Simulation study 1*

If between study variance was neglected, the false positive rate was 0.066 for the multivariate comparison, which is only slightly elevated compared to the required 0.05. If between study variance was estimated, the false positive rate was adequate, 0.050. For Bonferroni corrected univariate comparisons, the familywise false positive rate was 0.049 without estimated between study variance, and 0.047 with estimated between study variance. For uncorrected tests, the familywise error rate was too high; 0.306 without estimated between study variance, 0.276 with estimated between study variance.

3.5.2 *Simulation study 2*

If 0% of the data were missing, the false positive rate was 0.060 for the multivariate comparison. If 40% of the data were missing, it was 0.059, whereas it was 0.097 if 70% of the data were missing. For the uncorrected univariate comparisons, the familywise error rate was too high, around 0.3, for all three percentages missing. For the Bonferroni corrected univariate comparison, the familywise error rate was 0.046, 0.046, and 0.040 for 0%, 40% and 70% missing. The multivariate results show that false positive rate is not completely under control if the percentage missing test scores becomes very high.

With respect to power, as can be seen in Figure 3, uncorrected univariate comparisons show more significant results than multivariate or Bonferroni corrected univariate normative comparisons. Because familywise error was too high for uncorrected comparisons, the advantage in terms of power cannot be interpreted. Multivariate normative comparisons and Bonferroni corrected univariate comparisons show similar results in all conditions, with the exception of the 5 simulated deviations condition. When the patient deviates on 5 tests, the multivariate comparison is more sensitive.

Figure 3 also shows that sensitivity was about equal with 0% and 40% missing data. The comparisons with 70% missing data had slightly higher sensitivity. This should not be taken to suggest that 70% missing data is preferable, as the false positive rate was also slightly higher.

3.5.2.1 *Follow-up simulation studies*

As a follow-up, we investigated the effect of the magnitude of between study variance on the false positive rate. To this end we computed the intraclass correlation (ICC), which is defined as the ratio of the between study variance and the sum of the within and between study variance. In simulation studies 1 and 2 this ICC was 0.167. A preliminary analysis of the ANDI database shows ICCs ranging from 0 to 0.4, depending on the type of tests under study. These ICC's thus suggest that between study variance might vary considerably in real applications.

To investigate whether a larger between study variance affects false positive rate, we repeated simulation study 2 with a between study

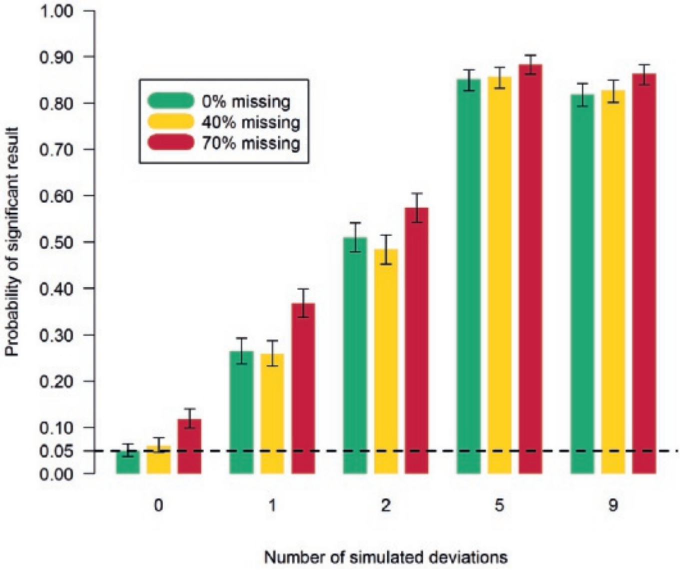

Figure 3.3: False positives (where number of deviations = 0) and sensitivity (where number of deviations > 0) as a function of the number of simulated deviations, for 0%, 40%, and 70% missing data in the norm group. Error bars represent 95% confidence intervals.

variance of 17, yielding an ICC of 0.4 (17 / (17 + 25)). With this higher level of between study variance, the false positive rates for multivariate normative comparisons were 0.060, 0.069 and 0.114 in the 0%, 40%, and 70% missing data conditions. For Bonferroni corrected univariate normative comparisons, the false positive rates for these conditions were 0.060, 0.066 and 0.074. For uncorrected univariate comparisons, the false positive rates were too high, around 0.35. These results indicate that false positive rate only slightly increases if between study variance increases.

In realistic settings, not every study will have the same number of participants, i.e. sample sizes will be unbalanced. To investigate its effects, we ran simulations with a mean of N=50, and a standard deviation of 10, with 70% missing data. These simulations showed a false positive rate of 0.112 for the multivariate comparisons, which is about the same as for the equal sample size case. Univariate uncorrected results showed a false positive rate of 0.302, while Bonferroni corrected results showed a false positive rate of 0.06. Therefore, unequal sample sizes do not seem to be problematic for multivariate or univariate comparisons. We also looked at simulations with unequal sample sizes and fewer participants, i.e a mean of N=25, and a standard deviation of 5. For these simulations, the multivariate comparisons showed a false positive rate of 0.192, while the univariate uncorrected result was 0.327 and the Bonferroni corrected result was 0.054. The false positive rate is increased for the multivariate result. This seems to be because the problems of 70% missing data are combined with a mean decrease of 50% of the number of participants in this condition.

All simulations so far have been done with ten tests. We also checked whether the same results were obtained for 20 and 5 tests. Fitting models to data from 20 tests took considerably more resources than fitting models with 10 tests. Therefore, we only ran the 70% missing condition, and performed 100 rather than 1000 simulations. With 5 tests, we ran 1000 simulations with a 60% missing condition, as 70% of 5 does not give a whole number of test scores to remove.

A total of 11 simulations with 20 tests showed convergence issues and had to be rerun, demonstrating that with more parameters, results can become more unstable with this amount of missing data. Multivariate results showed a false positive rate of 0.17. Uncorrected univariate results showed a false positive rate of 0.36. Bonferroni corrected results showed a false positive rate of 0.02. The elevated type 1 error rate for multivariate comparisons seems to originate in less precise estimates of covariances between tests: Because the number of participants and studies were kept equal, increasing the number of tests implies that the number of studies in which two tests are administered together decreases. So some covariance estimates are based on a single study with 30 participants. Because the Bonferroni corrected tests do not use covariance, they remain conservative.

With 5 tests and 60% missing, the false positive rate was 0.07 for the multivariate comparisons. Uncorrected univariate results showed a false positive rate of 0.214. Bonferroni corrected results showed a false positive rate of 0.055. This shows that with fewer tests, the multivariate method performs appropriately.

Lastly, we investigated the effect of including fewer studies, as fewer than 30 studies might be available for some neuropsychological tests. We simulated data with 20 studies for 10 tests with 40% missing data, because with 70% missing not all covariances could be estimated. The false positive rate was 0.07 for the multivariate method, 0.326 for the univariate uncorrected method and 0.055 for the Bonferroni corrected method. So although the number of studies, and therefore also participants, was cut by a third, false positives rates were not affected.

3.5.3 *Empirical example*

To give an impression of what the analysis would look like in practice, the method was applied to the ANDI database described earlier, and was used to examine the profile of a patient with Parkinson's disease. The details of the Parkinson's disease dataset have been described elsewhere (Muslimovic, Post, Speelman, & Schmand, 2005; Broeders et al., 2013).

Because the ANDI database contains many tests, we only selected tests that the patient had completed, and fitted the model to only those tests. For this example, the model was fitted to two variables of the Auditory Verbal Learning Test (AVLT), three variables of the Stroop test, two variables of the Trail Making Test (TMT), one variable of the Letter Fluency Test, one variable of the Semantic Fluency Test, summing up to a total of nine variables. For each of these variables, more than 1700 participants were available in the ANDI database (www.andi.nl/home). All variables were demographically corrected for age, sex and level of education, except for TMT part A, for which correction for sex was not necessary. All test variables were transformed to normality using Box-Cox transformations, and were recoded and standardized (de Vent et al., 2016).

In Figure 4, four bivariate plots are given for the patient with Parkinson's disease. A selection of two-dimensional plots is given because although the multivariate comparison provides a single result for eleven dimensions, this eleven-dimensional result is not easily visualized. As can be seen, correlations between variables differ, i.e. the shape of the bivariate distribution differs. The Stroop Color and Word variables in the top left plot are correlated, presumably because they belong to the same test and tap into the same naming speed component. The Stroop Color and TMT part b variables in the top right plot are only slightly correlated, presumably because although they both involve speed, one involves paper-and-pencil tracing, while the other involves verbal naming. Recalling words from memory after 30 minutes in the AVLT, and tracing a path in the TMT in the bottom left plot are completely uncorrelated, which is why the ellipse is circular. In all these bivariate plots, the patient falls within the 95% confidence interval. For the bottom right plot, this is not the case, as the patient falls far below the ellipse. This is mainly due to a very slow performance on the color-word interference condition of the Stroop. This slow performance is incongruent with the normal performance on the other Stroop subtask.

The multivariate test result is $T^2_{\text{norm}}(9, 30902) = 4.32$, $p < 0.001$. Using the one-sided criterion, we first have to ascertain whether the sum of differences is negative, which it is, -0.76. Therefore, we can conclude that this patient is impaired, as $p < 0.10$.

3.6 DISCUSSION

Multivariate normative comparisons are a valuable tool in neuropsychological assessment. Therefore, it is important that a multivariate normative database becomes available. We proposed the construction of such a multivariate database by joining healthy control group data

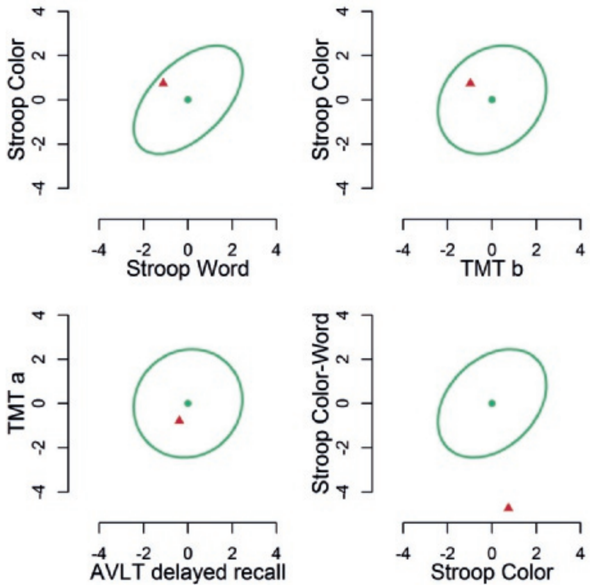

Figure 3.4: Four selected bivariate plots. The ellipses denote the 95% multivariate region. The dots denote the mean score in the norm group. The triangles depict the patient's scores.

from published neuropsychological studies. In this paper we also outlined a solution to three issues that arise when constructing such a combined database. First, test scores may differ between studies. Second, not all tests are administered in all studies. Third, patients should be compared to controls of a similar age, sex, and level of education. We developed a method that uses multilevel modeling to solve these three issues.

Our first set of simulations shows that estimating the variance between studies keeps false positive rate at an acceptable level. The results of our second set of simulations show that the number of false positives is too high if the percentage of missing data is 70%, but is satisfactory if 40% of the data is missing. Sensitivity of normative comparisons remains intact, even if 70% of the data is missing.

The power advantage, or enhanced sensitivity, of the multivariate comparison over Bonferroni corrected univariate comparisons was not visible in all conditions. Only when the patient deviated on half the tests, did the multivariate comparisons outperform Bonferroni corrected univariate comparisons. This is in line with earlier results (Huizenga et al., 2007), where it was shown that the advantage of the multivariate comparisons over univariate comparisons is greatest with intermediate numbers of deviations, with smaller advantages when the number of deviations is either very high or very low.

In the simulations with 20 tests and the simulations with smaller N, the false positive rate was not under control for the multivariate comparisons. This may be the result of the very large number of parameters needed in comparison to the number of participants, which primarily affected the estimates of the covariance between tests. A potential solution for such cases, if extra data collection is impossible, could be to provide restrictions on the covariances using a factor model, or to include prior information on the covariances.

Note that the proposed multilevel approach estimates between study variance. An alternative way to aggregate data over studies is to assume that between study variance does not need to be estimated. This assumption might in some applications be required if not sufficient studies are available to estimate this between study variance component (Hussong, Curran, & Bauer, 2013). Fortunately, in neuropsychology, sufficient studies are available as many studies administer the same instruments. Another alternative is to estimate between study variance, but to refrain from using it in comparisons. This may be more in line with current practice, where norms are used from a single normative study. However, we see the possibility to include between study variance as an advantage, as it allows for generalization, whereas assuming that between study variance is zero does not allow for generalization to new studies and new cases (Curran & Hussong, 2009).

The current approach requires several assumptions. First, the multilevel procedure assumes that all contributing studies have drawn random samples of healthy participants. At first sight, this assumption may not be met in neuropsychological studies. For example, some researchers will only draw random samples from one sex, e.g. women, because they are studying the effects of a particular disease that occurs predominantly in women, e.g. breast cancer. This matching will however be harmless to our assumption of random sampling, as the assumption pertains to the data after correction for age, sex, and educational background. As another example, close acquaintances of patients are popular controls: They are typically from similar educational backgrounds as the patient population and are often willing to participate (Gomez-Anson et al., 2007). Again, the fact that background is similar does not seem to be problematic, as educational background is included in the model. Finally, some control samples cannot be presupposed to be from the healthy population, such as non-schizophrenic psychiatric patients or even abstinent non-Korsakoff alcoholics (Moritz & Woodward, 2005; Oscar-Berman, Kirkley, Gansler, & Couture, 2004). These should not be included in the composite normative database.

Second, multilevel analysis assumes that the included studies are randomly sampled from a population of studies. In practice, all available studies that fit the inclusion criteria would be included, rather than taking a sample. Therefore, we argue that this assumption is likely to be met. Note that this is similar to a random-effects meta-analysis, where all studies, and not a random sample, on the effect under investigation are included.

Third, the current methodology may allow for missing data at the level of individual participants. This requires that the missing data mechanism can be considered ignorable. Fortunately, we do not expect many non-ignorable missing values in neuropsychological studies. Patients may find it difficult to complete test batteries, e.g. because of fatigue. Therefore, test batteries are designed to be not too demanding (Lezak et al., 2012). This implies that healthy participants often can complete the entire battery, and therefore few scores are generally missing. Because the number of non-ignorable missing data points, if present, should thus be small, the amount of bias in the estimates they incur will most likely be negligible.

Fourth, the normative comparison method assumes that scores are multivariate normally distributed around predicted scores. Little is known about the multivariate distribution of tests because large multivariate datasets have generally not been available. We do however know that violations of univariate normality, which preclude multivariate normality, are common in neuropsychology. Neuropsychological test scores may for example be skewed and truncated by ceiling or floor effects (Proust-Lima, Dartigues, & Jacqmin-Gadda, 2011). Statistical tests have been shown to be generally robust to mild violations of distributional assumptions in a group comparison setting (Jacqmin-Gadda et al., 2007) as well as in a normative comparison setting (Crawford et al., 2006) but more serious violations may result in a larger false positive rate. The multivariate comparison method has been shown to be robust to varying levels of skewness of the multivariate distribution but not to varying levels of kurtosis (Grasman et al., 2010). Solutions that have been proposed when multivariate normality is not tenable, involve transformations of the data (Looney, 1995) or non-parametric comparisons (Grasman et al., 2010).

Fifth, the current method requires calculation of the covariance between every pair of tests. Therefore, every test has to be administered with each of the other tests to at least a few participants. This limits the number of tests that can be included, as only the more common tests will have been administered together with all other tests. This was the case for the empirical example: A selection of tests had to be made to ensure that all covariances could be estimated with the present method. If less common tests need to be included, solutions may lie in models that restrict covariances, for example to obey a certain factor structure, or in collecting additional data (Carrig, Manrique-Vallier, Ranby, Reiter, & Hoyle, 2015).

The current approach can be extended in a variety of ways. First, although the proposed model flexibly handles missing data in test scores, it still resorts to listwise deletion of cases having a missing value on one of the covariates. Because missing covariates are handled differently from missing scores, this may result in many cases being dropped that were previously included. In these situations, alternatives to FIML such as multiple imputation might be a good solution.

This method can be extended beyond clinical neuropsychology, but note that clinical neuropsychology has three advantages that may not be present in every other field. The first is that neuropsychological test administration has been standardized to a high degree, such that data from different studies can be pooled. If there are for example differences between how tests are scored, additional steps may be necessary to harmonize measurements across studies (Hussong et al., 2013). The second advantage is that clinical neuropsychology is a large field, so many studies are available that have tested control groups. In smaller fields, it may be difficult to find sufficient studies that have administered the same test to accurately estimate between study variance. The third is that neuropsychologists administer multiple tests to the same participants, and therefore covariances between tests can be estimated. In fields where smaller test batteries are common, the lack of overlapping tests may imply that multivariate normative comparisons according to the current methodology are not feasible.

These advantages are however present in other fields, for example, in personnel psychology where highly standardized tests are regularly administered in large batteries. But also outside of psychology, the methods described here can be used just as easily for example in medicine, where physiological measures like blood pressure and heart rate are compared against the norm. Profiles of such measures could be compared against the norm as well using the multivariate method described here.

In conclusion, a large composite multivariate normative dataset can be established by combining data from many different studies. The current multilevel extension of multivariate normative comparisons can be used to handle (i) variability in test scores between studies (ii) missing data which arise because not all studies administered the same tests, and (iii) background variables. This multilevel extension allows routine multivariate comparisons of patients' test scores to multivariate normative data. This will enhance sensitivity of normative comparisons in neuropsychology, and may also be valuable in other contexts, e.g. in clinical or personnel psychology or medicine.

4

MULTIVARIATE NORMATIVE COMPARISONS FOR NEUROPSYCHOLOGICAL ASSESSMENT BY A MULTILEVEL FACTOR STRUCTURE OR MULTIPLE IMPUTATION APPROACH

4.1 ABSTRACT

Neuropsychologists administer neuropsychological tests to decide whether a patient is cognitively impaired. This clinical decision is made by comparing a patient's scores to those of healthy participants in a normative sample. In a multivariate normative comparison, a patient's entire profile of scores is compared to scores in a normative sample. Such a multivariate comparison has been shown to improve clinical decision making. However, it requires a multivariate normative dataset, which often is unavailable.

To obtain such a multivariate normative dataset, we propose to aggregate healthy control group data from existing neuropsychological studies. As not all studies administered the same tests, this aggregated database will contain substantial amounts of missing data. We therefore propose two solutions: multiple imputation and factor modeling.

Our simulation studies show that factor modeling is preferred over multiple imputation, provided that the factor model is adequately specified. This factor modeling approach will therefore allow routine use of multivariate normative comparisons, enabling more accurate clinical decision making.

4.2 INTRODUCTION

Normative comparisons are used to compare a patient's test scores to scores in a normative sample. In neuropsychological clinical practice, tests are designed to detect impairments in attention, working memory, inhibition or other cognitive functions. When normative comparisons show that a patient's scores are low compared to scores in a normative sample, this result may guide the treatment plan and can contribute to the characterization of the patient's condition which may be caused by a disease, like Alzheimer's disease or Parkinson's disease, or by brain damage due to traumatic injury or stroke (Lezak

[0] Published as: Agelink van Rentergem, J. A., de Vent, N. R., Schmand, B. A., Murre, J. M. J., & Huizenga, H. M. (2017). Multivariate normative comparisons for neuropsychological assessment by a multilevel factor structure or multiple imputation approach, *Psychological Assessment*. Advance online publication. doi:10.1037/pas0000489

et al., 2012; Strauss et al., 2006; Tierney et al., 1996). Normative comparisons are also used in neuropsychological research. They may be used to quantify the number of impaired scores in a treatment group as compared to a placebo group (Kraemer et al., 2003; e.g. Evans, Elliott, Reynders, & Isaac, 2014), or to assign participants to impaired or unimpaired groups. This grouping can then serve as an independent variable in studies investigating biomarkers or treatments (Meyer, Boscardin, Kwasa, & Price, 2013). As normative comparisons are ubiquitous in neuropsychological practice and research, it is important to optimize their performance.

Clinicians currently compare patient data to normative data collected by test publishers, who generally collect normative data for one test at a time. Normative data for a single test allow only univariate comparisons, in which a patient's score is compared to scores in a normative sample for each test separately (Crawford & Garthwaite, 2002; for applications, see for example Bird, Castelli, Malik, Frith, & Husain, 2004, or Cappelletti, Butterworth, & Kopelman, 2012).

In multivariate normative comparisons, all of a patient's test scores are simultaneously compared to those in the normative sample, to determine whether the profile of test scores is abnormal (Huizenga et al., 2007; Grasman et al., 2010; Huba, 1985; Crawford & Allan, 1994). One advantage is that they can identify deviating profiles that cannot be identified by multiple univariate comparisons. Deviating profiles may for example feature unexpected combinations of high scores on some tests, and low scores on others. Second, multivariate normative comparisons do not require corrections for multiple comparisons, as only a single comparison is made across tests (Huizenga et al., 2007). Therefore, one can perform this comparison without having to correct for an increased false positive rate due to multiple testing (Huizenga et al., 2016), and without having to estimate the number of univariate deviations one would expect in the healthy population (Brooks et al., 2009). Third, Su et al. (2015) showed that for research in HIV-related cognitive impairment, multivariate normative comparisons result in higher specificity than the univariate criteria that are commonly used. Multivariate normative comparisons have for example been used to study the psychological effects of Parkinson's disease, stroke and bacterial meningitis (Broeders et al., 2013; Castelli et al., 2010; Phaf et al., 2010; Schmand et al., 2010). Multivariate normative comparisons do not seem to have been broadly adopted in clinical practice. This may be caused by the unavailability of the required multivariate norm data.

Multivariate normative comparisons have many advantages, but require that multivariate normative data are available, i.e. that normative participants completed the same battery of tests that has been completed by the patient. However, clinicians and researchers draw from a variety of test batteries. For these ad hoc combinations of tests, multivariate normative data are generally unavailable. This limits the broader application of multivariate normative comparisons. In theory, this issue could be solved by administering all neuropsychological tests to one normative sample. However, because the total number of neuropsychological tests has become very large, administering all tests would be prohibitively expensive and taxing on participants constituting the healthy normative sample. Therefore, we develop a practical alternative in this article.

In a typical clinical neuropsychological study, multiple tests are administered to two groups: a clinical group of interest, and a control group that is healthy but is otherwise comparable to the clinical group. The control group data can be considered a small, but useful, multivariate normative dataset for a particular set of tests. If a neuropsychologist administers the same set of tests to a patient whose background characteristics are comparable to the control group, this clinician could use the control group data from this study to make a multivariate normative comparison.

However, a dataset from a single study is not useful to every clinician: Some tests that the clinician administers will not have been administered in the study, and the clinician's patient may not be comparable to the study participants. The clinician's patient may be younger than the participants in the study, or better educated. However, if the clinician would have access to a database consisting of multiple studies, chances increase that data are available for the clinician's tests. Also, if the database has many participants of different sexes, ages and levels of education, the clinician will be able to correct scores for the influence of these background variables. Therefore, it is useful to combine data from multiple studies to achieve a larger palette of tests, and to achieve better coverage of different sexes, ages and levels of education.

One data combination initiative in the field of neuropsychology is the Advanced Neuropsychological Diagnostics Infrastructure (ANDI; de Vent et al., 2016). As part of the ANDI project, neuropsychological test data of healthy participants have been aggregated into a single database. The database currently contains over twenty thousand participants from almost a hundred studies, with data for over thirty tests. For the most common neuropsychological tests that are frequently administered together, multivariate normative comparisons can be carried out using a multilevel approach (Agelink van Rentergem, Murre, & Huizenga, 2017). However, we will show that this is not the case for less common tests that are not often administered together, and we will propose and test two possible solutions.

The structure of this article is as follows. First, we describe issues that arise when combining control datasets from multiple studies in establishing a normative database. These are variability in scores between studies, scores that are missing because tests have not been administered in every study, and combinations of tests that have never been administered together in any of the studies. Second, we introduce two approaches which potentially solve all these issues, namely a multiple imputation and a factor structure approach. Third, we run simulation studies to test which of these approaches is most useful for normative comparisons. Fourth, we demonstrate the application of the factor structure approach to empirical data. Finally, we discuss potential limitations and improvements. We will continue with the use of clinical neuropsychology as a motivating example, although the methods described can be applied in any field where scores on highly standardized measures are compared to normative data, for example in personnel or clinical psychology

4.2.1 *Between study variance*

Scores on tests may differ from one healthy sample to the next because researchers' study design choices may affect scores. First, participants' motivation to achieve the best score may differ between studies. For example, in one study, an animal-naming test may be the very first test that participants have to complete, and they may be highly motivated to name as many animals as they can. In another study, they may have already completed an hour of other tests, and may be unmotivated on this animal-naming test and perform worse, even if test administration and the sampled population are identical (see Huizenga, van der Molen, Bexkens, Bos, & van den Wildenberg, 2012). Second, even though neuropsychological tests are standardized to a high degree, the way tests are administered can still differ between studies. For example, an experimenter may be required to call to attention any errors that the participant makes, but the kind of assistance offered and the speed at which mistakes are noticed and corrected can easily differ between experimenters (Snow, 1987, Lezak et al., 2012). Such kinds of between study differences can make participants' scores within studies more alike, and less like participants' scores from other studies. Therefore, a model that describes data from multiple sources ideally incorporates both variance between individuals as well as variance between studies.

Although presented here as an issue to be solved, the heterogeneity between studies can be viewed as a strength of the aggregated database. When a patient is compared to a single normative sample, the assumption is that the procedure and context are the same for both. This assumption may not be tenable, as variations on the intended procedure will occur in clinical practice just as they do in research. If enough studies can be sampled, the heterogeneity in test scores between studies, and thus across different contexts, can be estimated. Therefore, normative comparisons that take this source of variation into account may be more accurate than normative comparisons ignoring this source of variation.

4.2.2 *Structurally missing data*

When a test has not been administered in a study, this means that the score on this test is missing for all participants in that study.

Data with missing values can be analyzed using Full Information Maximum Likelihood (FIML; Graham et al., 2006). In FIML, each participant only contributes to the estimation of parameters involving the tests that the participant has completed. If a participant has completed two tests, the participant contributes to the estimation of only the means, variances and the covariance of those two tests. Whether FIML leads to correct estimates is dependent on the type of missing data, i.e. whether data are Missing Completely At Random, Missing At Random or Missing Not At Random (MCAR, MAR and MNAR; Schafer & Graham, 2002). These entail that the reason that data is missing is unrelated to the remainder of the data, is related to a known value in the data, or is related to a value that is unknown. If the type of missing data is one of the first two types, FIML will lead to correct parameter estimates.

Fortunately, in the current setting, healthy participants' test scores can be considered MCAR or at least MAR when the researcher has decided not to include a test in the study. One violation of MCAR could be that a researcher decides not to administer a test that has a strong ceiling effect in a young sample. However, because the variable age that explains these missing data is always recorded by both researchers and clinicians, the type of missing data is still MAR.

The issue of missing data thus might seem to be solved by FIML. Unfortunately, this is not the case. Missing data may complicate the estimation of covariance parameters between tests, if a combination of two tests has never been administered to a single participant. Missing combinations of tests therefore deserve separate attention, and will be discussed in the next paragraph.

4.2.3 *Missing combinations of tests*

Because the normative database is composed of studies that have already been conducted, there is no control over which tests are administered to whom. In some studies, memory will have been the primary focus, while other studies may focus on executive functions. Therefore, some tests will be administered together in many studies, and some tests will never be administered together. Chances are that such missing combinations of tests will always exist, even if a large number of studies are included.

A multivariate normative comparison uses the covariance between tests in determining whether a profile of scores is abnormal. If a combination of two tests has not been administered, calculating the covariance between these two tests is not straightforward. We here consider two potential solutions: using a multiple imputation approach, or a factor structure approach.

We first considered Multiple Imputation (MI). Many missing data problems can be handled by MI (Schafer & Graham, 2002; Rubin, 1986). MI uses a regression model to predict new values for those values that are missing, and adds simulated random error to these predicted values. This is done multiple times for the same variable, resulting in multiple complete versions of the same data that differ from each other due to the random error that was added. Because in the resulting complete datasets combinations of tests are no longer missing, all models can be fitted to these complete datasets. Note however that since some tests have never been administered together, we expect that MI does not yield adequate estimates of covariance between these tests, and thus will not be well-suited for normative comparisons.

The second option we considered is to use factor modeling, and calculate the covariances implied by the factor structure (Cudeck, 2000). If we assume that the covariance between test scores in the database arises from one or more underlying latent factors, we could calculate the implied covariance of two tests from their mutual dependence on these factors. Such an approach is feasible in the domain of neuropsychology, as numerous studies have shown that a factor structure can be used to describe covariances between neuropsychological tests (e.g. Greenaway, Smith, Tangalos, Geda, & Ivnik, 2009; Dowling, Hermann, La Rue, & Sager, 2010; Mitchell, Shaughnessy, Shirk, Yang, & Atri, 2012). If the factor model is accurately specified, i.e. the covariance between tests is indeed due to dependence on the same latent factor, the covariances should be accurately estimated and multivariate normative comparisons should be accurate as well.

To summarize, multivariate normative comparisons require multivariate data from a normative sample that is ideally diverse in terms of background variables. Such data can be obtained by constructing a composite normative database consisting of control data from published studies. This raises three issues. The first, between study variance, can be handled in a multilevel approach. The second, missing data, can be handled by estimating the model using FIML. The third, missing combinations of tests, can be handled either by switching from FIML to multiple imputation or by assuming a factor structure. We extend the multivariate normative comparisons method to accommodate these more complex models in the following sections, before comparing their performance in a simulation study.

4.3 METHOD

In this section, we first describe multivariate normative comparisons taking into account the effects of background variables. Second, between study variance is added to the comparison. Third, within study covariance with missing combinations of tests is obtained either by using MI, or by imposing a factor structure.

4.3.1 *Multivariate normative comparisons*

Normative comparisons are described for one patient i, that completed P tests ($p = 1, 2, ..., P$). This patient is compared to N healthy participants in the normative sample ($n = 1, 2, ..., N$), where each healthy participant participated in one of G studies ($g = 1, 2, ..., G$).

In a multivariate comparison, the patient's scores on several tests are compared simultaneously to scores of healthy participants (Huizenga et al., 2007, Crawford & Allan, 1994, Huba, 1985). This is achieved by adapting the multivariate Hotelling's T^2 statistic (Huizenga et al., 2007), yielding the following equation for the Multivariate Normative Comparison (MNC) statistic:

$$MNC\,statistic = \frac{N-P}{(N-1)P}\frac{1}{(N+1)/N}(y_i - \hat{y}_i)'S^{-1}(y_i - \hat{y}_i) \quad (4.1)$$

where y_i is a vector of length P containing the test scores of patient i, $\hat{y}_i$ is a vector of length P containing the normative predicted scores and S^-1 is the inverse of the covariance matrix of the tests in the normative sample, of size PxP. To evaluate whether a patient's profile of scores is abnormal, the MNC test statistic has to be referred to an F distribution with P and $N - P$ degrees of freedom (Huizenga et al., 2007). In equation 1, the patient's scores are compared to the patient's predicted scores given his or her background variables would he or she be a member of the healthy normative sample. These predicted scores $\hat{y}_i$ for a patient i equal:

$$\hat{y}_i = \begin{pmatrix} \hat{y}_{i1} \\ \hat{y}_{i2} \\ \vdots \\ \hat{y}_{iP} \end{pmatrix} = ([1, x_{1i}, x_{2i}, x_{3i}] \otimes I) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} \quad (4.2)$$

where x_{1i}, x_{2i} and x_{3i} are scores on background variables for patient i, which in neuropsychological settings often correspond to sex, age and level of education, $\otimes$ denotes the Kronecker-product, I is an identity matrix of size PxP, β_0 is a column vector of length P, containing the intercepts for every test. These intercepts can be interpreted as the test scores that are predicted if the values of sex, age and level of education are all equal to zero. β_1, β_2 and β_3 are column vectors of length P, which contain the effects of a background variable on each test. For example, β_1, gives the change in test score for each test that results from a one unit increase in background variable x_1.

4.3.2 *Modeling*

To get estimates of the regression coefficients in β (required in eq. 2), and variances and covariances in S (required in eq. 1), a model is fitted to the normative data. Because the variance in the normative data is due to between study and within study variability, the covariance matrix S as it appears in equation 1 is the sum of two covariance matrices: one for the between study residuals and one for the within study residuals. When we combine the model in equation 2 with between study residuals u_0, that are unique to every study g, and within study residuals ϵ, that are unique to every participant i, the full model becomes:

$$\hat{y}_{ig} = \begin{pmatrix} \hat{y}_{i1} \\ \hat{y}_{i2} \\ \vdots \\ \hat{y}_{iP} \end{pmatrix} = ([1, x_{1i}, x_{2i}, x_{3i}] \otimes I) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + u_{0g} + \epsilon_i \tag{4.3}$$

where u_{0g} is a column vector of length P containing the elements u_{0g1} to u_{0gP}, which refer to test-specific residuals unique to study g, ϵ_i is a column vector of length P containing the elements ϵ_{i1} to ϵ_{iP}, which refer to test-specific residuals unique to participant i. The covariance matrix S thus equals:

$$S = COV_{u0} + COV_{\epsilon} \tag{4.4}$$

COV_{u0} is a diagonal covariance matrix of size PxP of between study residual elements u_{0g1} to u_{0gP}. The diagonal elements of this matrix represent the between study variances of the different tests. Between studies, it is assumed that tests are uncorrelated, and therefore off-diagonal elements are zero. It is also assumed that the effects of the background variables do not differ between studies, which is why the β coefficients do not get a subscript.

COV_{ϵ} is a covariance matrix of size PxP of within study residual elements ϵ_{i1} to ϵ_{iP}. The elements on the diagonal of the covariance matrix represent the within study variances of the test scores. The off-diagonal elements represent the within study covariances between test scores. An example of the structure for the residual covariance matrix S, if we leave COV_{ϵ} completely unstructured, is given in Table 1.

Table 4.1: Unstructured Residual Covariance Matrix S for Four Tests

	test 1	test 2	test 3	test 4
test 1	$var(u_{01}) + var(\epsilon_1)$	$cov(\epsilon_2, \epsilon_1)$	$cov(\epsilon_3, \epsilon_1)$	$cov(\epsilon_4, \epsilon_1)$
test 2	$cov(\epsilon_2, \epsilon_1)$	$var(u_{02}) + var(\epsilon_2)$	$cov(\epsilon_3, \epsilon_2)$	$cov(\epsilon_4, \epsilon_2)$
test 3	$cov(\epsilon_3, \epsilon_1)$	$cov(\epsilon_3, \epsilon_2)$	$var(u_{03}) + var(\epsilon_3)$	$cov(\epsilon_4, \epsilon_3)$
test 4	$cov(\epsilon_4, \epsilon_1)$	$cov(\epsilon_4, \epsilon_2)$	$cov(\epsilon_4, \epsilon_3)$	$var(u_{04}) + var(\epsilon_4)$

An advantage of this unstructured approach is that every within study covariance is estimated freely, which allows for any pattern of correlations that may exist in the data. However, this model is not identified when combinations of tests are missing. If tests 1 and 4 are not administered to the same participants, no single value can be identified for $cov(\epsilon_4, \epsilon_1)$. However, it might be argued that we may still come to an estimate of $cov(\epsilon_4, \epsilon_1)$, if we use Multiple Imputation (MI).

4.3.3 *Multiple Imputation*

In MI, a model is fitted for every dependent variable with missing values, with other test scores and background variables as predictors. For each missing value, a predicted value is thus calculated and a random error term is generated. Together, these form a new plausible value. The prediction ensures that the imputed score is near to the scores of similar participants, and the random error term ensures that the amount of variance in the data does not decrease. The multiple imputation method runs this entire procedure multiple times. After the multiple imputation step is done, the model from equation 2 with an unstructured covariance matrix COV_ϵ is fitted to each imputed dataset. The average is computed over each of the fitted models for all the estimated parameters.

4.3.4 *Factor structure*

An alternative is to use a different specification of the within study covariance matrix COV_ϵ. If we assume that test scores are correlated because they are indicators of the same latent traits, we can estimate the latent factor structure, and calculate the covariances that are implied by this structure. Specifically, the covariance matrix COV_ϵ can be restricted to an M-factor structure, where M denotes the number of latent factors. In the M-factor model, the within study covariance matrix COV_ϵ is

$$COV_\epsilon = \Lambda\Psi\Lambda' + \Theta \tag{4.5}$$

where Λ is a factor loading matrix of size PxM relating P test scores to M latent factors, and Λ' is the transpose of this matrix. Ψ is a matrix of size MxM containing the variances of the latent factors on the diagonal, and the covariances between latent factors on the off-diagonal. Θ is a diagonal matrix of size PxP, containing the within study variances in test scores that are not explained by the latent factor or background variables. Θ is diagonal because given the latent factors, the tests should no longer be correlated (Bollen, 2002). If a test variable is included as an indicator for a latent factor, but is not correlated with the other test variables, the factor loading for this variable will be estimated to be 0. In this case, all variance in this test variable is not due to the latent factor, but is specific to this variable.

Cross-loadings, i.e. letting variables load on multiple factors, can be added to the matrix Λ. These cross-loadings will sometimes be necessary for the fit of the model. For example, if a test score on a memory scale is also determined by how well the participant comprehends the item verbally, a cross-loading with a verbal comprehension factor may be advisable. If this cross-loading is not included, the covariance with other verbal comprehension tasks may be underestimated. However, adding many cross-loadings makes the factor structure less stable, and should therefore be used only where necessary.

To identify the factor model, all factor variances on the diagonal of Ψ are set to 1. The elements on the off-diagonal of Ψ, the factor covariances, are freely estimated. Here we choose a confirmatory approach. That is, it is specified beforehand which variables load on which latent factor. In practice, one may need to explore multiple factor structure options that may be based on the literature (for example on Greenaway et al., 2009, or Mitchell et al., 2012), based on exploratory factor analysis (but see Fabrigar, Wegener, MacCallum, & Strahan, 1999), or based on a hybrid of the two, where multiple plausible models are compared using information criteria (Vrieze, 2012).

Substituting the within study covariance matrices in Equation 4, the complete residual covariance matrix S is modeled by:

$$S = COV_{u0} + \Lambda\Psi\Lambda' + \Theta \tag{4.6}$$

An example structure for S with one latent factor is given in Table 2.

Note that each off-diagonal element of the matrix in Table 2 is estimable, even if a combination of tests has not been administered to the same participants (Cudeck, 2000). Therefore, an important advantage of the factor structure approach is that it allows for estimation of covariance between tests that have not been administered together. Another advantage is that this gives a parsimonious description of the data, which may enhance the sensitivity of normative comparisons. A disadvantage of the factor model approach is that it assumes correct model specification.

Table 4.2: Residual Covariance Matrix S for Four Tests and One Latent Factor

	test 1	test 2	test 3	test 4
test 1	$var(u_{01}) + \lambda_{11}\psi_{11}\lambda_{11} + \theta_{11}$	$\lambda_{11}\psi_{11}\lambda_{21}$	$\lambda_{11}\psi_{11}\lambda_{31}$	$\lambda_{11}\psi_{11}\lambda_{41}$
test 2	$\lambda_{21}\psi_{11}\lambda_{11}$	$var(u_{02}) + \lambda_{21}\psi_{11}\lambda_{21} + \theta_{22}$	$\lambda_{21}\psi_{11}\lambda_{31}$	$\lambda_{21}\psi_{11}\lambda_{41}$
test 3	$\lambda_{31}\psi_{11}\lambda_{11}$	$\lambda_{31}\psi_{11}\lambda_{21}$	$var(u_{03}) + \lambda_{31}\psi_{11}\lambda_{31} + \theta_{33}$	$\lambda_{31}\psi_{11}\lambda_{41}$
test 4	$\lambda_{41}\psi_{11}\lambda_{11}$	$\lambda_{41}\psi_{11}\lambda_{21}$	$\lambda_{41}\psi_{11}\lambda_{31}$	$var(u_{04}) + \lambda_{41}\psi_{11}\lambda_{41} + \theta_{44}$

4.3.5 *Using the model estimates in comparisons*

After the model is fitted, either using multiple imputation or using a factor structure specification, the total covariance matrix of tests can be calculated. With this covariance matrix, normative comparisons can be made for a particular patient. Generally, the patient will have completed fewer tests than the normative sample, and only some elements of the estimated vectors and matrices will be relevant for the normative comparison. For the tests that the patient has completed, the relevant elements from the covariance matrix S are selected, and the relevant elements of the vectors β_0, β_1, β_2 and β_3 are selected to calculate the patient's predicted test scores $\hat{y}_i$ from eq. 1.

In applying the normative comparison method, the total number of participants in the norm group, N, figures in equation 2. What value to choose for N, which also impacts the degrees of freedom, is not straightforward, for two reasons. First, the observations are not independent due to the multilevel structure of the data. In what way this lack of independence between observations should be reflected in the choice of degrees of freedom is still subject of debate (Bolker et al., 2009). Second, the number of observations available per test can vary widely because some tests are not administered in as many studies as other tests, and because some studies are large-scale whereas others only investigated a few participants. These two factors leads to many possible choices of N. We have chosen the lowest number of observations on any of the tests in the comparison to be used as N (see for example Enders & Bandalos, 2001). This choice is conservative and thus reduces sensitivity. However, because studies are combined, the lowest number of observations will still be sizeable.

To summarize, normative comparisons require that we estimate regression coefficients, variances and covariances in the normative database. Because the normative database has a multilevel structure, the variance has to be estimated at two levels: a between study level and a within study level. The within study covariance between tests cannot be estimated in a straightforward manner if two tests have not been administered together, so a more elaborate approach is required. In the unstructured MI approach and the factor structure approach, we have two alternatives that each have their own advantages and disadvantages. The unstructured MI approach allows for any correlational pattern in the data. However, there is little information on covariance in the data when tests have not been administered together, and many covariance parameters have to be estimated in the unstructured approach. Therefore, covariance parameters may not be accurately estimated using the MI approach. The factor model can be used to estimate all covariance parameters. Further, it requires fewer parameters which may enhance sensitivity. However, the restrictive structure may not fit the correlational pattern in the data. For both methods, we test, in a simulation study, how well-behaved the normative comparisons are when many combinations of tests are missing.

4.4 SIMULATION STUDY

4.4.1 *Outcomes*

We examined whether the proposed procedures satisfy two requirements of normative comparisons. First, normative comparisons should control false positives (i.e. type 1 error): The proportion of comparisons that show deviations from the norm should be equal to prespecified levels (e.g. 0.05) for patients who in reality do not deviate. Second, the comparisons should have high sensitivity: Comparisons should be able to detect deviations that truly exist. All comparisons were two-sided, and 0.05 was used as the significance criterion for all comparisons.

4.4.2 *Parameter settings*

All specific parameter settings are given in the Supplementary Materials. The parameter settings were based on the documentation of the ANDI project (www.andi.nl, de Vent et al., 2016). This allowed for rough estimates of the size of the effects of background variables, the size of samples within studies, and the number of studies that would be contributed in an aggregate database of this type. The sample sizes and number of tests were however smaller in the simulations than in the ANDI database, as this speeded up computations in the simulations. Data were simulated for twelve tests.

4.4.3 *Simulation conditions*

The simulation conditions differed in (1) the factor model that was used to simulate data, (2) whether a patient was simulated to be different from the norm, (3) the pattern of missing data, and (4) the factor model that was fitted to the data.

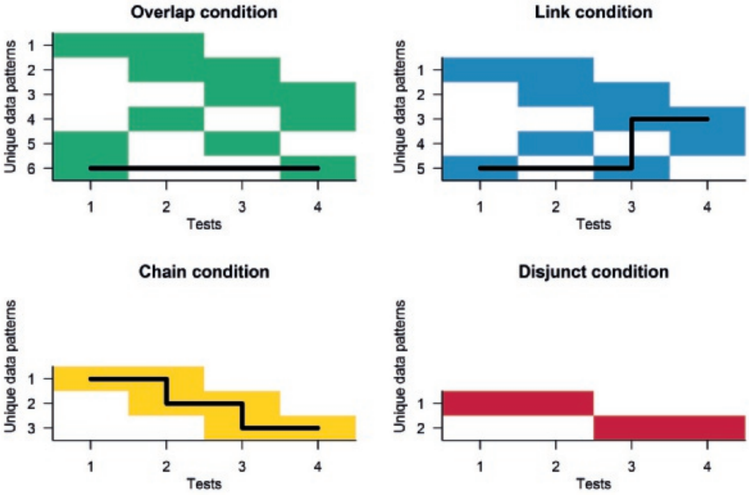

Figure 4.1: Unique missing data patterns for studies by simulated factor model, fitted model and missing data pattern. Colored boxes denote observed data, white boxes denote missing data. Black lines show how test 1 and test 4 are connected. This pattern is repeated three times for a total of 12 tests.

First, either a *one factor* or a *two factor* model was used to simulate normative data.

Second, the difference between the *false positive* condition and the *sensitivity* condition was introduced by manipulating the simulated patient data. In the *false positive* condition, the simulated patient's test scores were drawn from the same distributions as the normative data. In the *sensitivity* condition, a deviation on the first test was introduced by simulating this score from a distribution with a mean two standard deviations lower than the mean of healthy participants, where standard deviations were defined as the square roots of the diagonal of S (cf. eq 6).

Third, the missing pattern conditions were introduced by removing data points per study according to one of the patterns in Figure 1. Tests were a) administered together with each of the other tests in at least one study in the *overlap* condition, b) linked to other tests via other tests in the *link* condition, c) linked to other tests via a sequence of tests in the *chain* condition, or d) not linked via other tests in the *disjunct* condition. The percentage of missing data points was equal over conditions, i.e. 50%. In the simulated patient data, data were missing for tests 2, 3, 6, 7, 10, and 11.

Fourth, either an *unstructured*, a *one factor* or a *two factor* model was fitted to the data. The *unstructured* model was fitted using MI; the *one factor* and *two factor* models were fitted using FIML.

All combinations of models and data occurred, except that the *two factor model* was not fitted to the *one factor data*. For all conditions 1000 datasets were simulated. This allowed for sufficient precision in estimates of false positive rate and sensitivity. All models were fitted using *Mplus* and the *MplusAutomation* R-package (Muthén, 1997; Muthén, & Muthén, 2012; Hallquist, & Wiley, 2013).

Multiple imputations were performed using the *mice* R-package (van Buuren, & Groothuis-Oudshoorn, 2011). Because the variables that are used as predictors may have missing data as well, Multiple Imputation by Chained Equations (MICE; van Buuren, & Groothuis-Oudshoorn, 2011) uses variables with imputed values in the imputation of other variables. One simulated dataset is given as an example on github.com/JAvRZ/mplusRmodels, along with Mplus input to fit models, R code to extract the relevant parameters from the Mplus output files, and R code to perform multivariate normative comparisons given these parameters.

4.4.4 *Fitted factor model specification*

In the fitted *one factor* model, all factor loadings were estimated, i.e. all tests were indicators for the single latent factor. In the *two factor* model, to mimic a situation where the factor structure is already known from the literature, a model was fitted where each test loaded on the same factor as in the simulation, i.e. tests 1, 3, 5 etc. loaded on the first factor and tests 2, 4, 6 etc. loaded on the second factor. In the conditions where the simulated structure matched the fitted structure, the best possible performance of the factor models in the light of missing data could be evaluated. In the condition where data were simulated using a *two factor* model, and a *one factor* model was fitted, the consequences of misspecifying the factor structure could be evaluated.

4.4.5 *Multiple Imputation settings*

In the MICE procedure, a number of choices had to be made. First, a multilevel regression model with normally distributed errors was chosen to impute values, with within study variances assumed to be equal over studies. Multilevel models that allow within study variances to differ between studies show better imputation results (van Buuren, 2011), but require observations for every study for every test, which are not available for the current application. Second, background variables, and tests that were moderately or more highly correlated with the test to be imputed (i.e. $r \geq .10$, the default in the *mice* package, van Buuren, & Groothuis-Oudshoorn, 2011), were included as predictors in the imputation models. Third and fourth, the imputation algorithm ran for 10 iterations, and 50 complete datasets were

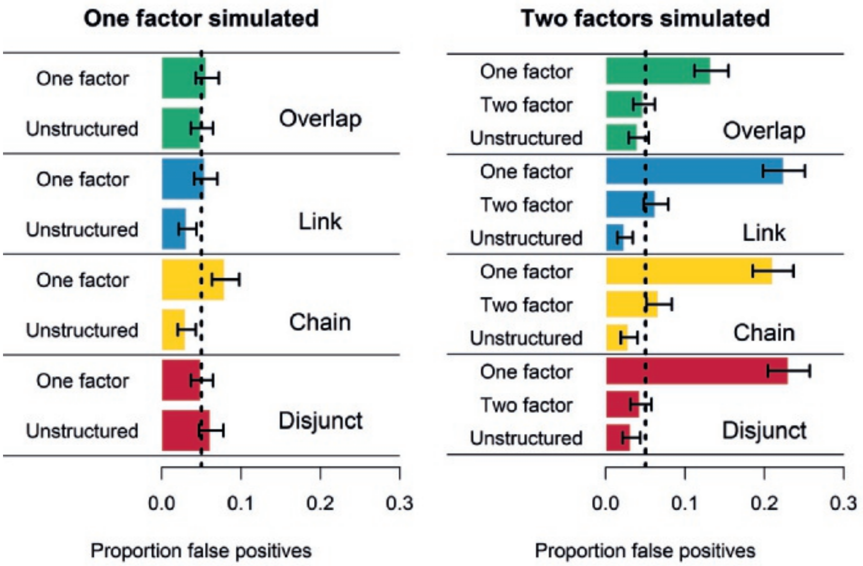

Figure 4.2: Barplot of false positive rate by simulated factor model, fitted model and missing data pattern. The black dashed line indicates the nominal false positive rate. Error bars represent 95% confidence intervals.

generated for each simulation (5 iterations and 5 imputed datasets are the default in the *mice* package).

4.5 RESULTS

The results in the *false positive* condition are presented in Figure 2. If *one factor* was simulated, irrespective of the pattern of missing data, the proportion of significant results was always close to the required .05 level. If *two factors* were simulated, irrespective of the pattern of missing data, the false positive rate with a *two factor* model or an *unstructured* model was always close to nominal. However, if the *one factor* model was fitted to *two factor* data, the false positive rate became unacceptably large, for all missing data patterns.

The results in the *sensitivity* condition are presented in Figure 3. In the *sensitivity* condition, if *one factor* was simulated, the *one factor* model outperformed the *unstructured* model. The advantage of the *one factor* model increased with the degree of missing information, as shown in Figure 3. The same holds true if *two factors* were simulated and modeled. Note however that the high sensitivity when fitting a *one factor* model to *two factor* data should be interpreted in the light of the elevated false positive rate and can therefore not be celebrated.

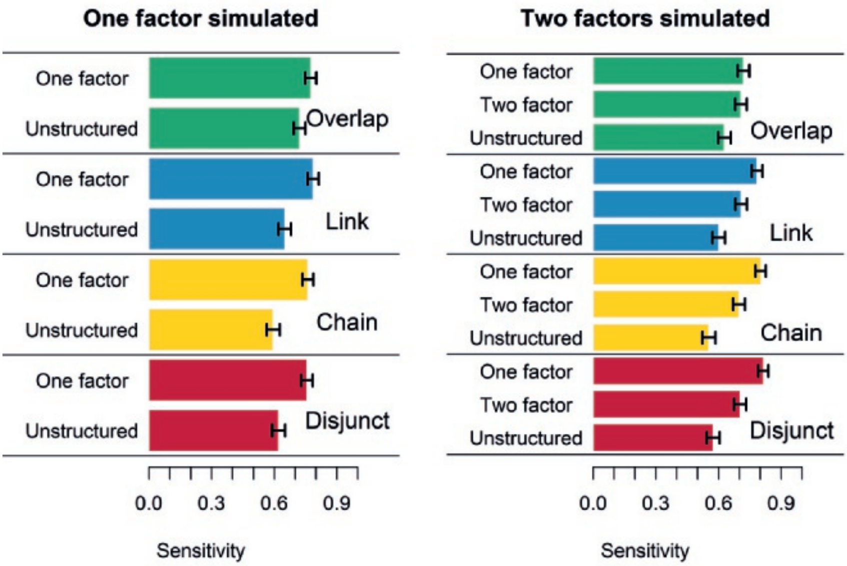

Figure 4.3: Barplot of sensitivity by simulated factor model, fitted model and missing data pattern. Error bars represent 95% confidence intervals.

We examined covariance parameters to investigate why sensitivity was higher for the *one factor* and *two factor* conditions than for the *unstructured* condition, and why the false positive rate was increased when fitting the *one factor* model to *two factor* data. Specifically, the covariance between tests 1 and 4 is of interest, as this is the parameter that should become more difficult to estimate when the extent of missing combinations worsens. These covariance estimates are plotted for every condition in Figure 4.

As shown in Figure 4, fitting the *one factor* model to *two factor* data leads to overestimates of the covariance, regardless of the pattern of missing data. This explains the increase in false positive rate observed in this condition. Fitting the *one factor* model to *one factor* data, and the *two factor* model to *two factor* data, led to correct estimation of covariances, regardless of the missing data pattern. For the *unstructured* condition, this was not the case, as the extent of the covariance underestimation increased when the combinations of test scores decreased. In the *unstructured-disjunct* condition, this covariance was even estimated as 0. This explains the drop in sensitivity that was observed between *factor* and *unstructured* conditions.

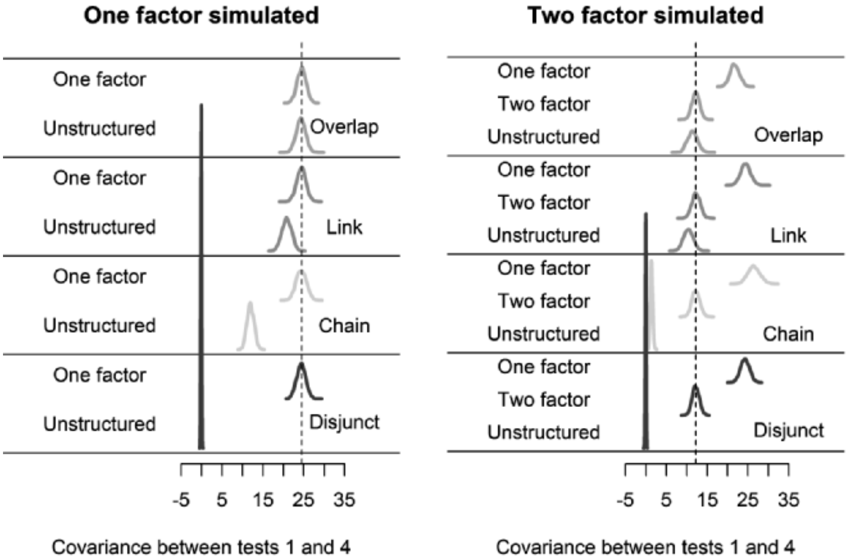

Figure 4.4: Density plot of covariance estimates by simulated factor model, fitted model and missing data pattern. The black dashed line indicates the covariance that was simulated. Note that in the unstructured disjunct condition all estimates were nearly zero, resulting in a very peaked distribution.

4.5.1 *Additional simulations*

So far, all latent factor models involved either one or two latent factors, to examine the behavior of the two missing data solutions in a controlled environment. In practice, more latent factors will be needed, as most neuropsychological assessments include more than two cognitive constructs. To investigate whether such a larger model, with more test variables and more parameters, could be fitted, we also performed simulations with a six-factor model. In these simulations, data on 24 variables were simulated, that each loaded on one of the six factors. The factor loadings were the same as in the previous simulations. In this factor model, we set all correlations between factors to 0.25. Missing values were introduced with the Link pattern from the previous simulations, again resulting in 50% missing data. Missing data were introduced to the patient data with the same pattern as in the previous simulations. No deviations were simulated. All 1000 simulations with a six-factor model converged. The false positive rate of the multivariate normative comparisons was 0.07, which is still close to nominal.

4.5.2 *Empirical example*

As an empirical example, we fit a factor model to a subset of the ANDI database (de Vent et al., 2016). The subset was selected to make sure that all variables were linked to all other variables, like in the Link condition in the simulations, and to guarantee that at least 100 participants were available for every variable. In total, 27 variables were selected from the WAIS III, Rey Complex Figure Task, Modified Wisconsin Card Sorting Test, Letter Fluency, Semantic Fluency, Trail Making Test, Stroop, Auditory Verbal Learning Test and the Boston Naming Test. All studies that contributed data to ANDI were approved by the research ethics committees of the institutions where the studies were conducted. The data have been checked for outliers, standardized, recoded, and transformed to normality, as is described elsewhere (de Vent et al., 2016).

A five factor model was fitted to the data, with the five factors representing Attention/Working Memory, Memory, Verbal Comprehension / Language, Executive Functions / Processing Speed and Perceptual Organisation. This factor model was constructed on the basis of models fitted in several neuropsychological papers (Pedraza et al., 2005, Dowling et al., 2010, Greenaway et al., 2009). However, it should not be considered definitive in any respect, and it is likely that a better model can be constructed. The model is represented in Figure 5.

In Mplus, the factor model was fitted, and the estimates of Λ, Ψ, Θ, COV_u0g and β were saved from the Mplus output. The lowest

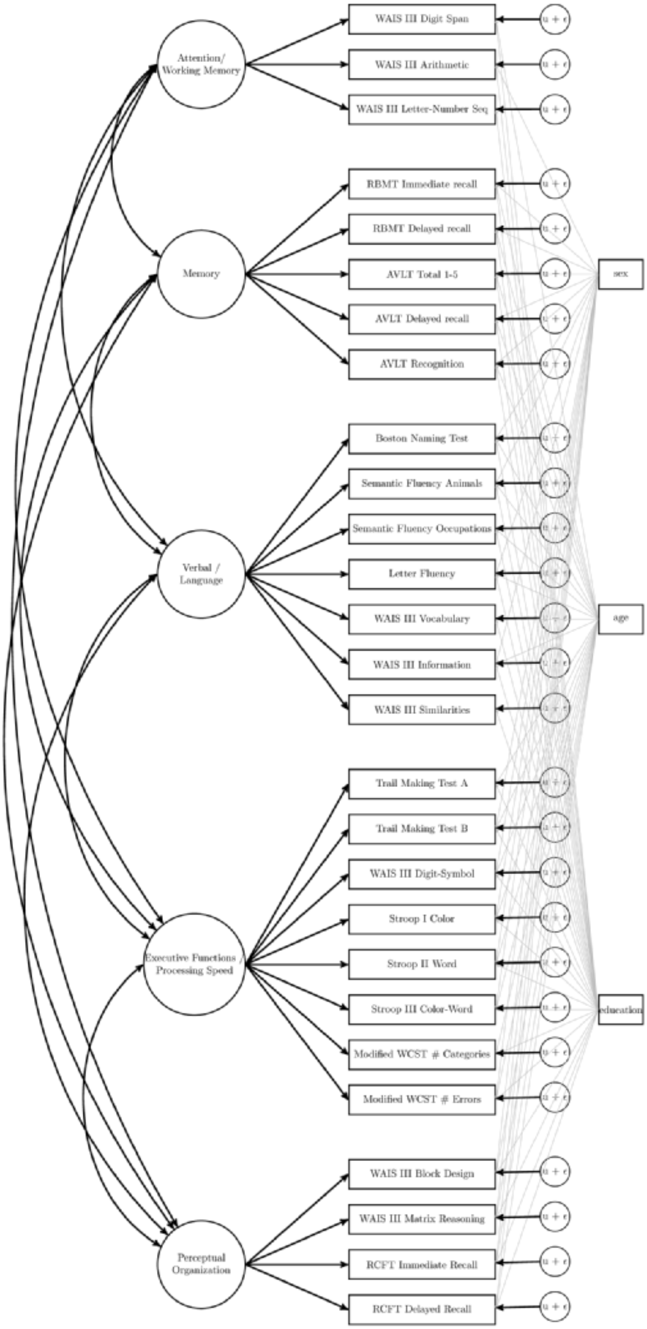

Figure 4.5: Example five factor model. Observed variables, i.e. test variables and demographic variables in rectangles. Unobserved variables, i.e. latent factors and error terms, in circles. Single-headed arrows, black and gray, denote effects and factor loadings. Double-headed arrows denote correlations. u and ϵ error terms are put in a single circle to simplify the representation, although the variance components are estimated separately.

N for these 27 variables was 153, so the numerator degrees of freedom would equal 27, and the denominator degrees of freedom would equal 126.

To demonstrate the procedure, we simulated data for a hypothetical highly educated female 76-year-old patient's data. The data are given in Table 3.

The multivariate normative comparison works as follows. First, the demographic variables and raw scores on the test variables are entered by the clinician. Second, the patient's raw test scores are transformed and standardized to be on the same scale as the transformed and standardized norm data (de Vent et al., 2016). Second, predicted values are computed for all test variables, using equation 2 with the demographic values of the patient, and the regression coefficients β. Third, the covariance matrix of these test variables is computed, using equation 6 with the estimates Λ, Ψ, Θ, COV_u0g. Fourth, the MNC statistic is computed, using equation 1, with the patient data, the predicted values, and the covariance matrix. The output of the analysis is given in Table 4.

Because the p-value in Table 4 is smaller than our threshold of 0.05, we can conclude that this patient deviates in a multivariate sense from the norm. Since this is a 27-dimensional result, it cannot be readily visualized. One option is to look at the profile of scores. These are presented in Figure 6. This however is not a truly multivariate presentation, as the correlations between tests are not visible.

Therefore, separate two-dimensional visualizations are also helpful. One of them is presented in Figure 7. In this Figure, the ellipse is very narrow because there is a high estimated correlation in the normative sample between RCFT Immediate Recall and RCFT Delayed Recall. The patient's combination of scores shows a deviation: The score on Immediate Recall is higher than predicted for a person of this sex, age, and level of education, while on Delayed Recall, the score is lower than predicted. Because these tests are so highly correlated, this combination of high and low scores is rare in the healthy population. A clinician could use this multivariate result to draw the conclusion that the patient's retention is worrying, given how typical the rest of the profile of test scores is.

4.6 DISCUSSION

Multivariate normative comparisons provide a new neuropsychological tool that is sensitive to subtle deviations in a patient's cognitive profile. However, these comparisons require that normative data are available for multiple tests from the same participants. We suggest to obtain the required normative data by providing a second life to data from control groups of neuropsychological studies, and aim to solve

Table 4.3: Simulated Scores on 27 Neuropsychological Test Variables.

Neuropsychological test variables	Score
Digit Span	18
Arithmetic	12
Letter-Number Sequencing	10
RBMT Immediate Recall	23
RBMT Delayed Recall	10
AVLT 1-5	50
AVLT Delayed Recall	5
AVLT Recognition	29
Boston Naming Test	50
Semantic Fluency Animals	27
Semantic Fluency Occupations	25
Letter Fluency	32
Vocabulary	36
Information	22
Similarities	24
TMT A	44
TMT B	90
Digit-Symbol	55
Stroop Word	45
Stroop Color	55
Stroop Color-Word	98
MWCST Categories	5
MWCST Errors	4
Block Design	40
Matrix Reasoning	21
RCFT Immediate Recall	23
RCFT Delayed Recall	5

Table 4.4: Output of the Multivariate Normative Comparison of the Simulated Scores from Table 3.

Patient ID	Sum of differences	Multivariate statistic	Degrees of freedom	p-value
1	-1.46	1.77	27, 126	0.02

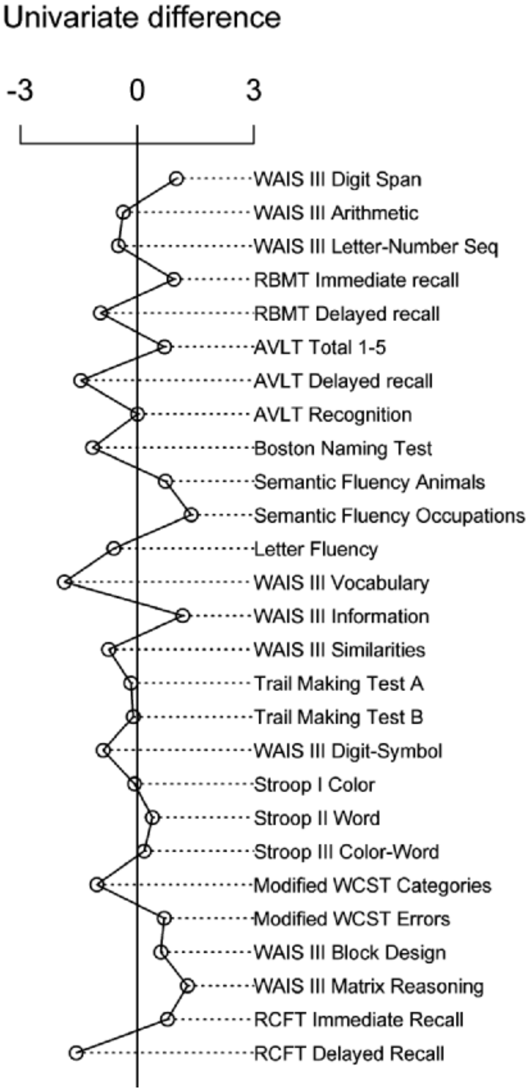

Figure 4.6: Line plot of the standardized difference between the simulated patient data in Table 3 and the values predicted for this patient.

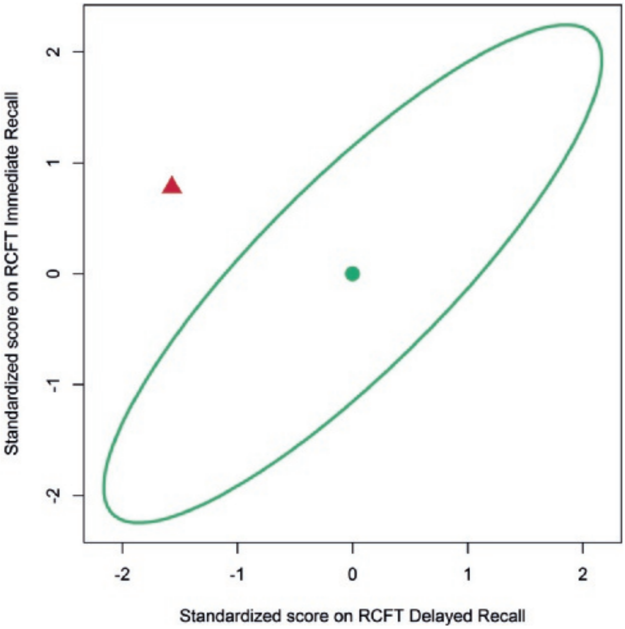

Figure 4.7: Example bivariate plot of the simulated patient data in Table 3, in comparison to the modeled normative data. The ellipse denotes 95% of the bivariate distribution. The triangle denotes the patient's combination of scores. The dot denotes the predicted score for a participant of the patient's sex, age and level of education.

several issues that arise when using such an aggregated neuropsychological database.

First, in combining data from multiple studies, random differences between studies need to be accounted for. This issue is solved by a multilevel approach in which the variance between studies is estimated. Second, if different tests are administered in different studies, scores are practically missing for whole studies. This issue is solved by using estimation methods that do not require complete data for every test. Third, if tests have never been administered together in any of the studies, the covariance between tests cannot be estimated. In this article, two alternatives were considered that may potentially solve this last issue, namely a multiple imputation method and a factor structure method.

In a simulation study, the performance of the factor structure and multiple imputation methods was evaluated according to two criteria: whether the false positive rate of the multivariate comparisons was appropriate, and whether their sensitivity to detect true abnormalities was high. The simulations show that false positive rate is adequate for both multiple imputation and factor modeling, although the latter only if the factor structure is adequately specified. With respect to sensitivity, simulations show that factor modeling outperforms multiple imputation. We therefore conclude that when the factor structure is adequately specified, the factor structure method is preferred and that when this is not the case, the multiple imputation method is the method of choice.

The proposed procedure rests on a number of assumptions. First, it is assumed that the within study variance is homoscedastic. For example, it is assumed that the variance in test scores is equal between healthy controls with a low and high education. This assumption could very well be tenable for a wide variety of tests, but it may be violated for tests that show a ceiling effect in the younger population, or that require rapid responding. Variance may then be larger for older age groups than for younger age groups (Rabbitt, 1979). If this assumption is indeed violated, the model could be extended to allow heteroscedastic variances. A patient's score can then be evaluated using a normative comparison that includes a variance term appropriate for the patient's sex, age and education.

Second, it is assumed that the within study covariance is also equal among different levels of the background variables. Again, in the eventuality that this assumption is violated, the model could be extended to allow differences in covariances. The appropriate covariance for the patient's background would then also be included in the normative comparison.

Third, it is assumed that the within study variances and covariances are equal over different studies, as there is no reason for variances of tests or covariances between tests to differ between studies.

However, in applications in which within study variances and covariances would differ between studies, this assumption can be relaxed.

Fourth, it is assumed that the effects of background variables are equal over different studies, and thus that there are no random effects of these background variables. Note that restriction of the range of the background variables within studies may lead to different effects in different studies. For example, studies that only administer tests to students may show no effects of age on scores, while studies that administer tests across the lifespan may show large effects of age. Although such random effects can be added to the model, we prefer not to include them as they are artificially introduced by restriction of range.

Fifth, it is assumed in several steps of the procedure, both in the multiple imputation as well as in the factor structure method, that the residuals are normally distributed for every test. A solution to violations would be to transform scores, which is already common practice in neuropsychology (Jacqmin-Gadda, Sibillot, Proust, Molina, & Thiébaut, 2007). Not all variables lend themselves to transformations to normality. For some tests, the skew will be so drastic that no transformation will result in a normal distribution. To give an example, some tasks involve a recognition trial, where the participant has to recognize stimuli that were used in the task they have just performed. In clinical samples, these variables can be clinically relevant, as not recognizing all stimuli may be indicative of memory impairment. In healthy samples however, these variables have no variance, as practically all participants obtain the maximum score.

Because of the parametric assumptions of the procedure, it would be difficult to include such variables into the multivariate comparison. However, it might also not be worthwhile to do so. First, variables that have no variance in the healthy population will also be uncorrelated with all other tests in the healthy population. Therefore, the multivariate procedure does not contribute anything beyond univariate comparisons for these variables. Second, these variables may not require a statistical assessment, as any score that is below the optimal score is presumably a red flag to a clinician, even before normative data are consulted.

Sixth, it is assumed for the factor model that the same factor structure holds for different studies. This assumption should be established empirically. Dowling et al. (2010) found that their five-factor model of neuropsychological tests was invariant across different age groups and sexes, which is promising, but does not ensure that this is the case across studies. Also, it is assumed that the factor structure is stable across sexes and across all ages and education levels that are included in the sample. This assumption may be violated in the very old (e.g. 85 and older), as those participants' cognitive function may be better described by a lower number of factors than younger individuals, due to a process known as dedifferentiation (Li et al, 2004). The collapse into fewer cognitive factors might imply that tests are more highly correlated in the very old than in other age groups. As a result, sensitivity could be artificially reduced in this older age group if the correlation structure of the whole age range is used.

Seventh, the factor model assumes that if the latent factors are taken into account, the test scores are uncorrelated. In general however, some test scores will remain correlated, even after the factor is taken into account. For example, two variables can have a higher correlation than is expected from their dependence on the same latent memory factor, because the two variables are both measured after a delay of 30 minutes. To account for such an extra dependency, two solutions are available. First, if enough variables of this type exist in the dataset, a new "delayed recall factor" may be inferred, that explains the additional correlation between these variables. Second, if data is available on the correlation between two delayed recall measures, this additional correlation may be added to the model directly, by estimating the corresponding off-diagonal element of the residual covariance matrix Θ, which was previously constrained to 0. Adding these additional residual correlations to the Θ-matrix might be a good solution, but adding too many of these reduces the stability of the original factor model.

It is important to note that all assumptions noted above are made about the distribution of the data in the normative sample, not the patient sample. Therefore, distributions are allowed to be very different in clinical populations. For example, a different factor model holds for patients with Alzheimer's disease than for healthy participants (Siedlecki, Honig, & Stern, 2008). As another example, some test variables will in a clinical population not become normally distributed after transformation. For the multivariate normative comparisons procedure, this is not an issue, as the assumptions pertain to the normative sample.

Aside from assumptions, several considerations merit attention. First, neither the multiple imputation method nor the factor structure method is an automatic procedure that can be applied without further thought. For the multiple imputation method, it should be checked that multiple imputation has succeeded, for example, that imputations produce realistic values for every variable (van Buuren & Groothuis-Oudshoorn, 2011). For the factor structure method, the appropriateness of the model should be assessed using goodness-of-fit measures (Schermelleh-Engel, Moosbrugger, & Müller, 2003) and ideally a validation dataset. For the purposes of the normative comparison, the goal of the factor model is not so much to obtain the true factor model underlying the data. Rather, the goal is to recover the covariance matrix as appropriately as possible. Therefore, different factor structures that generate nearly equivalent covariance matrices will give rise to similar normative comparisons. Nevertheless, we intend to perform a meta-analysis of factor structures proposed in the literature, to arrive at a factor model that approximates the covariance matrix well for the most common neuropsychological tests.

Second, we propose to first estimate the full covariance matrix for all tests in the database, and then select the relevant elements for the tests that the patient at hand has completed. An alternative would be to estimate a model for just the tests that the patient has completed. We prefer the current approach because of the beneficial nature of adding auxiliary tests in estimation (Enders, 2006; Graham, 2003; Graham, 2009) and the computational burden of fitting a separate model for every new patient.

Third, the variance between studies was included in the normative comparisons with good reason, as this variance might include differences in scores that arises from between study differences in motivation and administration. Methods that are currently in use do not account for such between study variances, because they use a single normative dataset. Therefore, patients are compared to less variable scores, which leads to higher sensitivity to deviations. Although the new method is appropriately modeling between study variance, sensitivity may be lower because of it. To get this sensitivity back, the patient could also be compared to the within study covariance matrix instead of the full covariance matrix, to get results that are more consistent with current practice, at the cost of decreased specificity.

Fourth, collecting data, fitting models and applying normative comparisons is not feasible for the typical user in clinical practice. To help the user, the methods that have been described in this article are currently being implemented in an interactive website. This website will allow clinical neuropsychologists to make multivariate comparisons on a daily basis using the ANDI database (de Vent et al., 2016).

Fifth, the quality of normative comparisons is reliant on the quality of the normative data. If the normative data contains a sample of highly motivated volunteer participants that outperform participants that do not want to participate, this biases the normative comparisons. Bias may be reduced because we include demographic corrections: If educated individuals are more likely to participate in a particular study, this does not affect our estimates because we correct for the influence of education. Bias may also be reduced because we include variance between studies: If non-representative samples are included in some but not all studies, their influence is diminished, because the model allows for variation between studies.

However, the analysis still rests on the assumption that in general, the included samples are similar to the general population. Therefore, in assembling a database as described here, datasets should be selected with this ulterior goal in mind. Samples that consist of friends and family of patients are probably similar to the general population, and community samples are generally collected with the goal of representativeness in mind. But perhaps, studies with convenience samples that actively volunteered for this type of study should be excluded to not bias the normative sample.

Normative comparisons can be impacted by non-random missing data in the normative dataset. For example, if the very old are typically unable to complete a particular task, the very old participants that do not have missing data on this task may be unrepresentative of the very old population. Therefore, the data should be scrutinized for tests that show selective missing data in sensitive populations (e.g. the very old and lower education levels).

One additional check that can be applied would be to compare the univariate distributions of test scores within the database, to distributions of scores that are reported in test manuals. If the univariate distributions are similar for all variables, this would provide evidence that the database is unbiased.

Finally, the current methodology can easily be extended to other domains. Normative comparisons are common in many fields of psychology where one is interested in the performance of an individual, for example in personnel, educational or clinical psychology, and in medicine. In those fields, a similar data combination venture could be undertaken, as control group data are abundant in those research fields as well.

In conclusion, the methods proposed in this article enable multivariate normative comparisons that could not be made before due to lack of multivariate normative data. If the factor structure can be adequately specified, the factor modeling approach should be preferred. If not, an unstructured multiple imputation approach is the method of choice. In this way, without requiring any new data collection, multivariate normative comparisons can be used as a sensitive tool to aid clinical decision making in science and in clinical practice.

COGNITIVE DOMAINS IN NEUROPSYCHOLOGY: SUPPORT FOR THE CATTELL-HORN-CARROLL MODEL IN TWO RESEARCH SYNTHESES

5.1 ABSTRACT

Many neuropsychologists are of the opinion that the multitude of cognitive tests may be grouped into a much smaller number of cognitive domains. However, there is little consensus on how many domains exist, what these domains are, nor on which cognitive tests belong to which domain. This incertitude can be solved by factor analysis, provided that the analysis includes a broad range of cognitive tests that have been administered to a very large number of people. In this article, two such factor analyses were performed, each combining multiple studies. The first analysis was a factor meta-analysis of correlation matrices, combining data of 60,398 healthy participants from 52 studies. Several models from the literature were fitted, of which a relatively complex model, based on the Cattell-Horn-Carroll (CHC) model, was found to describe the correlations much better than the others. The second analysis was a factor analysis of the Advanced Neuropsychological Diagnostics Infrastructure (ANDI) database, combining scores of 11,881 participants from 54 Dutch and Belgian studies not included in the first meta-analysis. Again, the model fit was much better for the CHC model than for the other models. Therefore, we conclude that the CHC model best describes which cognitive domains there are and which test belongs to which domain. Therefore, although it was originally developed in the intelligence literature, it deserves more attention in neuropsychology.

5.2 INTRODUCTION

Neuropsychological tests are designed to measure cognitive functions, which may be impaired by brain disorders like Alzheimer's or Parkinson's disease, traumatic brain injury, or stroke. The tests neuropsychologists use are often assigned to cognitive domains, such as executive function, memory, or attention.

There are many reasons for establishing domains of cognitive functions, and for assigning tests to these domains. The first reason may

[0] Submitted as: Agelink van Rentergem, J. A., de Vent, N. R., Schmand, B. A., Murre, J. M. J., Staaks, J. P. C., ANDI Consortium & Huizenga, H. M. (2017). *Cognitive domains in neuropsychology: Support for the Cattell-Horn-Carroll model in two research syntheses*. Manuscript submitted for publication.

be that a clinician suspects problems in a specific cognitive domain for a particular patient, and wants to select tests from this domain to administer. For example, if a patient comes in with subjective memory complaints, memory could be investigated further by selecting tests from this domain. The second reason may be that a clinician wants to qualify whether a particular patient is suffering from impairment on a single domain or on multiple domains. In the literature on mild cognitive impairment (MCI), for example, single-domain or multi-domain MCI are considered separate entities, which have separate prognoses (Petersen, 2004). The third reason may be that a clinician or researcher wants to use composite scores on cognitive domains as an outcome measure, rather than separate test scores. This method can reduce noise from individual measurement instruments (but see Lezak et al., 2012). These composite scores may be calculated by summing the scores of individual tests that belong to a particular domain, as is done in the calculation of performance IQ or verbal IQ. A more sophisticated approach is to obtain estimates of a latent trait through factor analysis or item response theory analysis of a single domain, and use scores on the latent trait as an outcome measure (Gross et al., 2015). The fourth reason may be to establish the validity of a particular test. If a researcher designs a new test intended to measure memory, he or she can calculate whether scores correlate highly with other tests in the memory domain, and do not correlate as highly with tests from other domains. Therefore, domains can be used to show both convergent and divergent validity.

Although domains have many uses, the idea of domains of cognitive functions is not without problems. There is a lack of consensus on which tests belong to which domain, because there are many reasonable ways to assign tests to domains. For example, the Trail Making Test B (TMT B), in which one has to draw a line from labeled circles 1 to A to 2 to B to 3 etc., is one test that is particularly difficult to assign. Because it involves drawing with a pencil, and the outcome measure is the time to completion, one could assign it to the domain of psychomotor speed, along with tests like pegboard tests of manual dexterity. However, because TMT B performance depends for a large part on how attentive the person is, one could assign it to the domain of attention as well, along with tests like the Continuous Performance Test. Moreover, because it involves shifting back and forth between letters and numbers, one could assign it to the domain of executive functions, along with the Stroop Interference test.

There is also a lack of consensus on how many domains there are. For example, there are many tests that aim to assess memory in neuropsychology. Whether a single memory domain is sufficient, or whether more domains are necessary, is a matter of debate (Delis, Jacobson, Bondi, Hamilton, & Salmon, 2003). Measures of memory could be divided into measures of an immediate recall domain and a delayed recall domain, or in measures of a visuospatial memory domain and a verbal memory domain. Of course, one could also argue that separate domains are necessary for immediate visuospatial recall and delayed visuospatial recall.

A factor analysis can provide some clarity through quantification of what model best describes the correlations between tests. However, the resulting domains depend on the method and sample of the study, as we will outline next.

First, the factor structure that is found can depend on the tests that are selected. For example, if a test like TMT B is administered together with tests that measure executive functioning, TMT B may also load on a single executive functioning factor because it has elements of shifting. However, if more speeded tests are administered, TMT B may load on a different latent factor, processing speed, together with other measures of processing speed. Therefore, the domain to which a test seems to belong is dependent on the battery of tests used. Consequently, comparisons across studies with different batteries of tests become necessary.

Second, age can affect the factor structure that is found in a study, because age affects scores on almost all neuropsychological measures. Therefore, in a sample with a large age range, variables may become correlated because they depend on the same age variable. Elderly people generally score lower on all variables, and young people generally score high on all variables. If age is not appropriately accounted for, fitting a factor model to a sample with a large age range can provide support for a single "cognitive" factor, on which some participants score poorly - the elderly - and others score well - the young. One solution would be to study the factor structure in a sample that is homogeneous in age. However, since studying a single age group limits generalizability, an appropriate alternative is to include age in the analysis.

Third, and similarly to the age range effect, there can be a confounding effect of level of education in factor analysis. There is generally a large effect of education on neuropsychological test scores. Again, this may lead to the conclusion that to explain correlations between tests, we need just a single "cognitive" factor, on which some participants score poorly - those with little education - and some score well - those with much education. Such a single factor due to education would not be found in samples with very similar educational background, such as college students. However, since neuropsychological test results need to generalize beyond groups such as college students, it may again be more appropriate to correct for the effect of education in the analysis.

Fourth, domains can be different depending on the sample used. This is especially true for samples of patients with very specific deficits. A delayed recall test can become uncorrelated with other memory tests if delayed memory specifically is impaired by disorder or injury. Therefore, the structure of domains is ideally studied separately for healthy groups and different clinical groups. Results so far have shown that the factor structure has large communalities for many different clinical groups (Bowden, Cook, Bardenhagen, Shores, & Carstairs, 2004; Park et al., 2012; Schretlen et al., 2013), but it cannot be assumed that this is the case for all disorders.

Fifth, to get stable results for a factor analysis, many participants have to be tested on multiple tests. The amount of variance that is explained by latent factors may be low in neuropsychology, while there may be many latent factors, which increases the required sample size (MacCallum, Widaman, Zhang, & Hong, 1999). However, obtaining a large sample size for a battery of neuropsychological tests is costly, as the tests require training to administer, are administered one-on-one, and are time-consuming. This limits the number of participants that can be tested in a study, or limits the size of the battery that can be administered to a large number of participants.

Our goal is to establish how neuropsychological tests should be assigned to domains. We will do so by using a factor analytic approach, comparing different factor models that have been formulated in the literature. We will use the results of multiple studies, so we can achieve a broad range of neuropsychological tests, and we will correct for effects of demographic variables like age and level of education. We will study healthy adults, so the factor models are not confounded by sample differences in clinical status. Last, through combining different studies, samples of participants are combined to arrive at a much larger sample size than possible with a single study.

First, we will perform a factor analysis of neuropsychological tests, by applying a meta-analytic framework that allows for structural equation models to be fitted to summary statistics (Cheung & Chan, 2005). Specifically, this method pools correlation matrices from multiple studies to arrive at a single correlation matrix. To this correlation matrix, multiple models can be fitted, which allows us to compare the fit of neuropsychological factor models that have been formulated in the literature. Second, we will conduct a factor analysis of data from the Advanced Neuropsychological Diagnostics Infrastructure (ANDI) normative database (de Vent et al., 2016a). This database contains raw data from healthy control participants from multiple studies conducted in the Netherlands and Belgium, not included in our first analysis.

To summarize, neuropsychology would benefit from clarity on the number and type of cognitive domains, and on which tests belong to which cognitive domains. This would facilitate test selection, diagnosis of single-domain and multi-domain disorders, calculation of composite scores, and neuropsychological research into the construct validity of tests.

5.3 STUDY 1: FACTOR META-ANALYSIS

5.3.1 *Methods*

5.3.1.1 *Literature search*

A systematic literature search was conducted using PsycINFO and MEDLINE for articles that contained a factor analysis of neuropsychological tests in healthy adults. Factor analyses were chosen, as studies conducting a factor analysis generally recruit a large sample and administer a large battery of tests. The search strategy was developed in PsycINFO (see Appendix 1 for the syntax), because PsycINFO is particularly well-suited for searching psychological tests. The search strategy for MEDLINE was based on the PsycINFO search strategy. The search strategy consisted of the following key concepts: factor analysis-related terms, specific neuropsychological test-related terms and general neuropsychology-related terms. Deduplication of results was done using *Refworks*, and screening of results for inclusion was done using *Rayyan* (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016).

5.3.1.2 *Exclusion criteria*

The goal was to obtain for each article a healthy adult sample correlation matrix, containing both neuropsychological tests and demographic variables. Articles were excluded if a) fewer than two tests of interest were used, b) an adult sample was not studied, c) they were published before 1997, d) a typical sample was not studied, e) test administration was manipulated or otherwise differed from typical administration, f) they were included in the ANDI database. Criterion c was chosen because datasets published twenty years before the literature search could not be expected to still be available from the original authors. Criterion d entailed that we did not include groups with psychiatric or neurological disorders (e.g., bipolar disorder or epilepsy), with disorders that could interfere with test administration (e.g., hearing loss), or with conditions that were studied for their cognitive implications (e.g., HIV). Criterion e excluded studies in which manipulations (e.g., TMS) were applied to participants during testing, or in which novel, often computerized, versions of test batteries were used. This last choice was made because these novel versions are less familiar and less thoroughly validated than the common versions. Criterion f preserved the independence of the analyses done in study 1 and study 2 of the present article.

5.3.1.3 *Tests*

A complete list of variables that were considered of interest is given in Appendix 2. However, not every combination of variables was present in the correlation matrices that were analyzed. For twelve test variables, correlations were available with every other variable. To increase the number of usable correlations, different versions of the same test variables were combined (see Table 1). These tests may not be completely parallel, as there may be differences in test administration and scoring rules. However, the current analysis assumes that, although there may be mean differences between versions, the correlations with other test variables will not be different. This issue is addressed in study 2.

5.3.1.4 *Contacting authors*

With a few exceptions (e.g. Adrover-Roig, Sesé, Barceló, & Palmer, 2012), articles and/or supplementary materials did not contain the correlation matrix including both the tests and the demographic variables that were necessary for this study. Therefore, corresponding authors of included studies were contacted. In case a researcher appeared multiple times as a corresponding author in the included studies, a single, recent article was chosen which included a large selection of tests. In this case, if the corresponding authors agreed to share a correlation matrix, they were asked whether they would be willing to share the correlation matrix for other articles as well. If authors did not respond, they were reminded after a period of 2-3 weeks.

The authors were sent a list of variables of interest that were to be included, which were the test variables that they collected in their study, along with age, sex, and level of education. There was no specific hypothesis for the influence of sex on the factor structure, but we chose to correct for its influence as well because this is common in neuropsychology (Testa, Winicki, Pearlson, Gordon, & Schretlen, 2009). Level of education was scored differently in different studies, sometimes using a seven-point-scale, sometimes using years of education. This issue is discussed in more depth in the discussion section and is addressed in study 2. Authors were requested to send a correlation matrix of these variables, for the cognitively healthy sample within their data. If they were unsure that their participants qualified as cognitively healthy, possibilities for exclusion criteria within their data were discussed. For example, if measurements from the Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975) and Clinical Dementia Rating (CDR; Morris, 1997) had been taken in their study, participants with MMSE scores below 24 and CDR scores above 0 could be removed before the correlation matrix was computed. Since these exclusion criteria depended on what the authors had available in their data, this was an ad-hoc procedure.

Table 5.1: Included Test Variables.

Test variable	Abbreviation	Additional information
Trail Making Test Part A	TMTA	Combined with Color Trails Test Part 1, D-KEFS Trail Making Test condition 2.
Trail Making Test Part B	TMTB	Combined with Color Trails Test Part 2, D-KEFS Trail Making Test condition 4.
Logical Memory I	LMI	Combined across multiple WMS versions, combined with RBANS Story Immediate Memory.
Logical Memory II	LMII	Combined across multiple WMS versions, combined with RBANS Story Delayed Memory.
Letter Fluency	LF	Synonyms: Controlled Oral Word Association Test, Phonemic Verbal Fluency.
Semantic Fluency	SF	Synonyms: Categorical Verbal Fluency. Preferential inclusion of the "Animals" version if multiple were available.
Digit Span Forwards	DSF	Combined across multiple WAIS and WMS versions.
Digit Span Backwards	DSB	Combined across multiple WAIS and WMS versions.
Coding	COD	Combined across multiple WAIS versions. Synonym: Digit Symbol Substitution.
Boston Naming Test	BNT	
Auditory Verbal Learning Test – Total Recall	VLT-TR	Combined with California Verbal Learning Test – Total Recall, the Hopkins Verbal Learning Test – Total Recall, and RBANS List Learning.
Auditory Verbal Learning Test – Delayed Recall	VLT-DR	Combined with California Verbal Learning Test – Long-Delay Recall, the Hopkins Verbal Learning Test – Delayed Recall, and RBANS List Recall.

5.3.1.5 *Analysis*

The analysis was carried out using *R* (R Core Team, 2016). First, for each study, the correlation matrix was converted to a partial correlation matrix by partialing out the influence of age, sex, and level of education, using the *psych* package (Revelle, 2010).

A factor meta-analysis of the partial correlation matrices was conducted using the *metaSEM* package (Cheung, 2014). This factor meta-analysis consisted of two steps in itself (Cheung & Chan, 2005, Jak, 2015). First, the partial correlation matrices were pooled into a single weighted partial correlation matrix, using the total number of participants after exclusion for each study in the weighting. Second, using the weighted partial correlation matrix as input, different factor models that have been described in the literature were compared. For each model, fit was evaluated by χ^2, RMSEA, SRMR, CFI, AIC, and BIC, using the rules of thumb outlined in Schermelleh-Engel et al. (2003) to decide what constitutes bad, acceptable and good fit.

5.3.1.6 *Candidate factor models*

Factor models that were broad enough to span all neuropsychological tests were selected from the literature. This excludes factor models that describe correlations between tests from just a single domain (e.g. Huizinga, Dolan, & van der Molen, 2006). The first model was a model with a single latent factor on which all variables loaded. Verhaeghen and Salthouse (1997) used a single factor model in a meta-analysis of correlations of neuropsychological test scores, and found that a large part of the variance in test scores can be construed as variance on a single common latent factor. The fit of the one factor model can be used as a reference to judge the fit of more complex models.

The second and third models came from the chapter structure of the clinical neuropsychology reference works by Strauss et al. (2006) and Lezak et al. (2012). Although there is not an explicit factor model in these works, the neuropsychological tests are categorized into separate chapters. Therefore, they give a good impression of which tests belong together in the eyes of clinical neuropsychologists. In Strauss et al. (2006), the chapters containing the included tests were "General cognitive functioning", "Executive Functions", "Memory", "Orientation and attention" and "Language". In Lezak et al. (2012), the chapters containing the included tests were "Attention", "Memory", "Executive Functions", "Verbal functions and language skills". The difference between the two was that Digit Span and Coding fall under "General cognitive functioning" in Strauss et al. (2006), and under "Orientation and attention" in Lezak et al. (2012).

The fourth and fifth models were based on the opinion of experts. The fourth model was based on the domains used in Gross et al. (2015). Gross et al. (2015) assigned tests to "Memory", "Executive functioning" and “Rest” domains on the basis of expert opinion. Of the currently included tests, only the Boston Naming Test fell in the "Rest" category. The fifth model was based on a survey of clinical neuropsychologists (Hoogland et al., 2017). Twenty experts were asked to rate, on a seven-point Likert scale, how well test variables assess cognitive functioning on a particular domain. For the twelve tests included here, the relevant domains were "Language", "Attention and working memory", "Memory" and "Executive function". For the factor model used, all mean ratings were above 4.85 on the seven-point scale, indicating a large degree of confidence that these variables should be assigned to these domains.

The sixth model was based on the recommendations made by Larrabee (2014). Larrabee (2014) divided tests in six domains, on the basis of a review of the literature. This domain specification was explicitly intended to help clinicians compose a battery of tests that assesses cognitive abilities from different domains. The four domains for the included tests are "Verbal symbolic abilities", "Attention/working memory", "Processing speed", and "Learning and memory—verbal and visual".

The seventh and eight models were two variants of the Cattell-Horn-Carroll (CHC) model as described by Jewsbury et al. (2016). The CHC model was developed in intelligence research, rather than in clinical neuropsychology (McGrew, 2009). Jewsbury et al. (2016) demonstrated that the CHC model fits well in each of the nine neuropsychological datasets they studied, with only minor adaptations for each dataset. The factors for the included tests were the same across the two variants of the CHC model: "Acquired knowledge or crystallized ability", "Processing speed", "Long-term memory encoding and retrieval", "Working memory", and "Word fluency". In the first variant, TMTB measures "Processing speed". In the second variant, TMTB measures both "Processing speed" and "Working memory". All factor model specifications are given in Table 2.

Each factor model consisted of the following elements, which were freely estimated: Factor loadings describing the relationship between the tests and the latent variables, residual variances of the test variables, and covariances between latent variables. The covariances between latent variables can be interpreted as correlations, because all latent variable variances were fixed to 1.

5.3.2 *Results*

5.3.2.1 *Sample*

From the literature search, 3259 sources were identified. After deduplication, 2520 distinct sources remained. These were judged against the exclusion criteria, by inspection of the title, abstract, and description of the tests and measures that is provided in PsycINFO. After

Table 5.2: Factor Model Specifications of the Candidate Models for Study 1. Tests that Load on the Same Latent Factor Share a Letter. Some Tests Load on Multiple Latent Factors in the Hoogland and Jewsbury Models.

	TMTA	TMTB	LMI	LMII	LF	SF	DSF	DSB	COD	BNT	VLT-TR	VLT-DR
One factor	A	A	A	A	A	A	A	A	A	A	A	A
Strauss	D	D	C	C	B	B	A	A	A	E	C	C
Lezak	A	A	B	B	C	C	A	A	A	D	B	B
Gross	A	A	B	B	A	A	A	A	A	C	B	B
Hoogland	B	B + D	C	C	A + D	A + D	B	B	B	A	C	C
Larrabee	A	A	B	B	C	C	D	D	A	C	B	B
Jewsbury 1	B	B	A + C	A + C	E	E	D	D	B	A	C	C
Jewsbury 2	B	B + D	A + C	A + C	E	E	D	D	B	A	C	C

Note: TMTA = Trail Making Test A, TMTB = Trail Making Test B, LMI = Logical Memory I, LMII = Logical Memory II, LF = Letter Fluency, SF = Semantic Fluency, DSF = Digit Span Forwards, DSB = Digit Span Backwards, COD = Digit Symbol Substitution / Coding, BNT = Boston Naming Test, VLT-TR = Verbal Learning Test - Total Recall, VLT-DR = Verbal Learning Test - Delayed Recall.

this step, 330 articles were selected, of which the full-texts were obtained. Seven articles were excluded because the full-text could not be retrieved, so a total of 323 were eligible for inclusion. After e-mailing the corresponding authors, 60 correlation matrices were obtained from 57 studies (Adrover-Roig et al., 2012; Andrejeva et al., 2016; Andreotti & Hawkins; 2015; Albert et al., 2010; Barnes et al., 2016; Bennett & Stark, 2016; Bezdicek et al., 2014; Booth et al., 2015; Bouazzaoui et al., 2013; Bowden et al., 2004; Bunce, Batterham, Christensen, & Mackinnon, 2014; Burns, Nettelbeck, & McPherson, 2009; Chan et al., 2009; Chen et al., 2017; Ciccarelli et al., 2012; Darst et al., 2015; DeYoung, Peterson, & Higgins, 2005; Duff et al., 2006; Eifler et al., 2014; Ferreira et al., 2015; Fernaeus, Östberg, Wahlund, & Hellström, 2014; Fortin & Caza, 2014; Gallagher, Gray, Watson, Young, Ferrier, 2014; Hedden & Yoon, 2006; Hedden et al., 2014; Horvat et al., 2014; Hueng et al., 2011; Kafadar, 2012; Karagiannopoulou et al., 2016; Kesse-Guyot, Andreeva, Lassale, Hercberg, & Galan, 2014; Kim et al., 2013; Komulainen et al., 2008; Krueger, Wilson, Bennett, & Aggarwal, 2009; Laukka et al., 2013; Lehrner et al., 2014; Liebel et al., 2017, Llinàs-Reglà et al., 2017; Mohn, Lystad, Ueland, Falkum, & Rund, 2017; Morrens et al., 2008; Ojeda et al., 2012; de Paula et al., 2013; Reppermund et al., 2011; Ricarte et al., 2016; Royall, Bishnoi, & Palmer, 2015; Schmidt et al., 2017, Siedlecki et al., 2010; Snitz et al., 2015; Sternäng, Lövdén, Kabir, Hamadani, & Wahlin, 2016; Thibeau, McFall, Wiebe, Anstey, & Dixon, 2016; Tractenberg et al., 2010; Tse, Balota, Yap, Duchek, & McCabe, 2010; Tuokko et al., 2009; Valenzuela & Sachdev, 2007; Waldinger, Cohen, Schulz, & Cromwell, 2015; Watts, Loskutova, Burns, & Johnson, 2013; Wettstein, Kuźma, Wahl, & Heyl, 2016; Williams, Suchy, & Kraybill, 2010). Horvat et al. (2014) provided four separate correlation matrices from four countries.

From these correlation matrices, tests were selected that were administered together in multiple studies. This limited the number of tests to the twelve described in the methods section. Five studies did not include any or just one of the selected tests, and were not included in the final analysis (Thibeau et al., 2016; Sternäng et al., 2016; DeYoung et al., 2005; Kafadar, 2012; Burns et al., 2009). The PRISMA diagram is given in Figure 1.

All correlations of test variables were scrutinized for miscoding. One source showed aberrant correlations that could not be explained: TMT B was positively correlated with other, unspeeded, tests in one correlation matrix (the oddity of which was noted in the original publication; Royall et al., 2015). Correlations with the TMT B variable were removed for this study. Motivating plots for this removal are provided in Appendix 3, along with the analysis which did include these correlations.

The final sample consisted of 60,398 participants and 55 correlation matrices. Study characteristics are given in Appendix 4, along with those correlation matrices for which we received explicit permission to share them here (49 out of 55). The correlations with age, sex, and level of education were partialed out from each correlation matrix. Variance in correlations between studies could not be estimated because for some pairs of tests only a few studies were available. Therefore a fixed-effects rather than a random-effects model was used to arrive at the pooled partial correlation matrix. The pooled partial correlation matrix is given in Table 3.

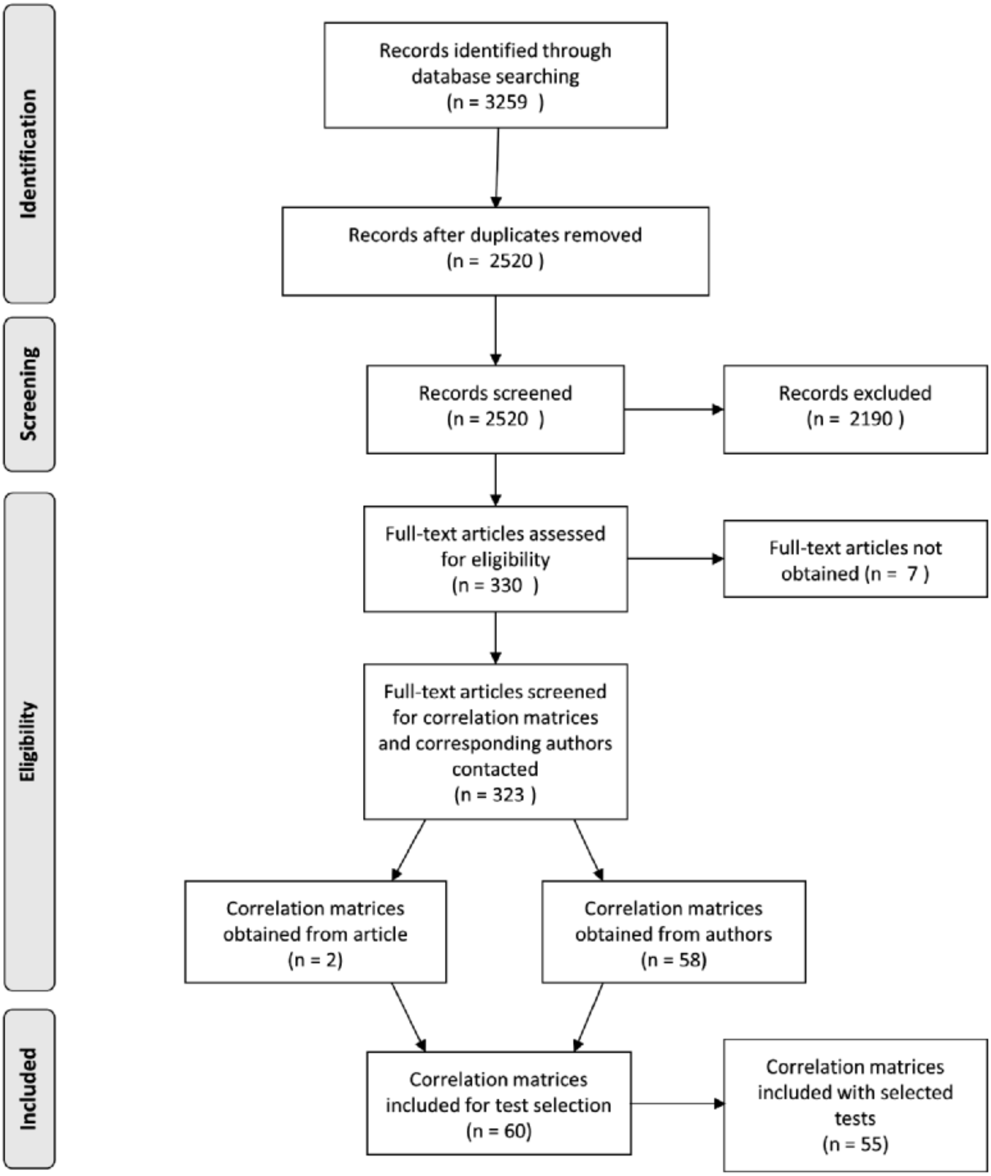
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 3259)
Records after duplicates removed (n = 2520)
Records screened (n = 2520)
Records excluded (n = 2190)
Full-text articles assessed for eligibility (n = 330)
Full-text articles not obtained (n = 7)
Full-text articles screened for correlation matrices and corresponding authors contacted (n = 323)
Correlation matrices obtained from article (n = 2)
Correlation matrices obtained from authors (n = 58)
Correlation matrices included for test selection (n = 60)
Correlation matrices included with selected tests (n = 55)

Figure 5.1: PRISMA diagram.

Table 5.3: Pooled Partial Correlation Matrix. In Parentheses (Italics): Standard Errors of Estimated Mean Correlations According to the Fixed Effects Model.

	TMTA	TMTB	LMI	LMII	LF	SF	DSF	DSB	COD	BNT	VLT-TR
TMTB	0.543 *(0.006)*										
LMI	-0.084 *(0.012)*	-0.171 *(0.011)*									
LMII	-0.091 *(0.012)*	-0.176 *(0.011)*	0.864 *(0.002)*								
LF	-0.207 *(0.009)*	-0.264 *(0.009)*	0.198 *(0.015)*	0.208 *(0.014)*							
SF	-0.222 *(0.009)*	-0.256 *(0.008)*	0.262 *(0.009)*	0.274 *(0.009)*	0.457 *(0.006)*						
DSF	-0.117 *(0.010)*	-0.202 *(0.010)*	0.151 *(0.010)*	0.134 *(0.010)*	0.231 *(0.011)*	0.168 *(0.009)*					
DSB	-0.159 *(0.011)*	-0.283 *(0.010)*	0.236 *(0.010)*	0.220 *(0.011)*	0.272 *(0.011)*	0.230 *(0.009)*	0.481 *(0.007)*				
COD	-0.487 *(0.012)*	-0.516 *(0.012)*	0.239 *(0.011)*	0.256 *(0.011)*	0.351 *(0.012)*	0.348 *(0.008)*	0.187 *(0.010)*	0.271 *(0.010)*			
BNT	-0.185 *(0.013)*	-0.195 *(0.012)*	0.258 *(0.010)*	0.267 *(0.010)*	0.234 *(0.011)*	0.284 *(0.008)*	0.128 *(0.012)*	0.153 *(0.014)*	0.304 *(0.012)*		
VLT-TR	-0.154 *(0.017)*	-0.217 *(0.018)*	0.447 *(0.010)*	0.461 (0.009)	0.267 *(0.018)*	0.349 *(0.004)*	0.196 *(0.011)*	0.271 *(0.012)*	0.300 *(0.011)*	0.232 *(0.010)*	
VLT-DR	-0.140 *(0.018)*	-0.154 *(0.018)*	0.439 *(0.010)*	0.502 (0.009)	0.197 *(0.022)*	0.322 *(0.008)*	0.092 *(0.012)*	0.183 *(0.012)*	0.278 *(0.012)*	0.225 *(0.010)*	0.695 *(0.005)*

Note. TMTA = Trail Making Test A, TMTB = Trail Making Test B, LMI = Logical Memory I, LMII = Logical Memory II, LF = Letter Fluency, SF = Semantic Fluency, DSF = Digit Span Forwards, DSB = Digit Span Backwards, COD = Digit Symbol Substitution / Coding, BNT = Boston Naming Test, VLT-TR = Verbal Learning Test - Total Recall, VLT-DR = Verbal Learning Test - Delayed Recall.

Table 5.4: Model Comparison Results.

	χ^2 (df)	RMSEA	SRMR	CFI	AIC	BIC
One factor	10411.2 (54)	0.056	0.218	0.941	10303.2	9816.8
Gross	6186.2 (51)	0.045	0.145	0.965	6084.2	5624.8
Hoogland*	4522.0 (45)	0.041	0.118	0.975	4432.0	4026.6
Lezak*	4635.7 (48)	0.040	0.121	0.974	4539.7	4107.3
Strauss*	3785.3 (44)	0.038	0.112	0.979	3697.3	3300.9
Larrabee	2831.5 (48)	0.031	0.098	0.984	2735.5	2303.1
Jewsbury 1	1334.0 (42)	0.023	0.060	0.993	1250.0	871.7
Jewsbury 2	1289.5 (41)	0.022	0.060	0.993	1207.5	838.2

*Model did not converge.

5.3.2.2 *Model fit*

The results of the model comparison between candidate models is given in Table 4. The Hoogland et al. (2017), Lezak et al. (2012), and Strauss et al. (2006) models did not converge. Therefore the fit measures for these models should be interpreted with caution. With respect to relative fit, the AIC and BIC indicate that the two variants of the complex Jewsbury model fit better than the other models.

With respect to absolute fit, the fit measures generally agree about the ordering of the models as well. All χ^2 values indicate bad fit (all χ^2 / df > 3), which suggests that none of the models provides exact fit. All RMSEA values indicate good fit (all RMSEA < 0.05), except for the one factor model, where the RMSEA value indicates acceptable fit (RMSEA < 0.08). The SRMR values indicate bad fit for the five simplest models (SRMR > 0.10), and acceptable fit for the models described by Larrabee and Jewsbury et al. (SRMR > 0.05). The CFI values indicate bad fit for the one factor model (CFI < 0.95), acceptable fit for the model used by Gross et al. (CFI < 0.97), and good fit for the other models (CFI > 0.97).

The best-fitting Jewsbury model is depicted in Figure 2, in which correlations between latent variables are also provided. Because Trail Making Test A and B are measured in time to completion, these variables and the "Processing Speed" factor that they loaded on, are reverse coded. Therefore, the negative correlations between "Processing Speed" and the other latent factors should be interpreted such that better "Processing Speed" is correlated with better scores on the other latent factors.

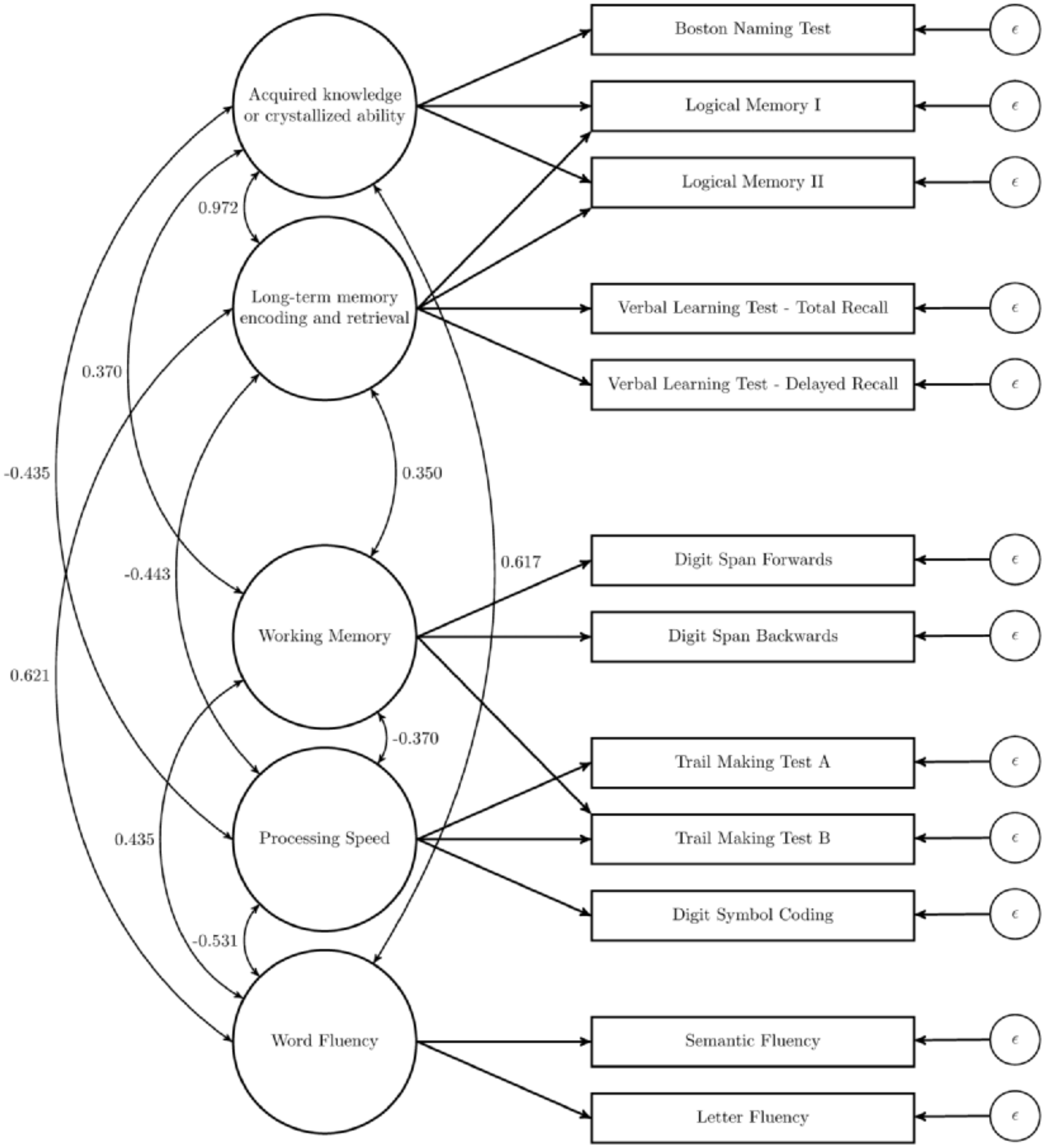

Figure 5.2: Jewsbury 2 model for the twelve tests included in study 1. For each combination of latent factors, the correlation is given.

5.3.3 *Discussion*

From this factor meta-analysis, we can conclude that the two Jewsbury models provide the best fit. This is remarkable, because AIC and BIC fit measures penalize complexity, and these two models are the most complex. The two Jewsbury models themselves do not differ by much, but all fit measures agree that the second model, with the extra cross-loading, fits better. Therefore, we conclude that for the tests used here, the correlations between test variables are best described by five cognitive domains, namely "Acquired knowledge or crystallized ability", "Processing speed", "Long-term memory encoding and retrieval", "Working memory", and "Word fluency". We also conclude that some test variables load on multiple of these domains.

The factor meta-analysis framework has several advantages, in that it allows for the analysis of a large number of tests and a very large number of participants. Using the partial correlation matrices rather than the raw correlation matrices allowed us to correct for the effects of age, sex, and level of education.

However, there are a number of limitations to this analysis. First, different versions of tests were used as if they are parallel. For example, correlations with the Hopkins Verbal Learning Test, California Verbal Learning Test, Rey Auditory Verbal Learning Test and Word List Recall of the RBANS were treated as if these versions are identical. This choice was made to arrive at a greater degree of test overlap between studies. However, there are differences between test versions in test administration, the number of repetitions, and the number of words that need to be remembered. The assumption here was that the correlations between the sum score variable and other test variables does not change due to these differences. This assumption may not be tenable.

Second, there were differences in education scales and education systems between studies. As argued in the introduction, it is necessary to remove the confounding influence of education. However, the contributing studies used different ways of coding level of education, which means that the correction in the form of the partial correlation was different between studies as well. Also, even if two studies used the same scale such as years of education, such a scale may have a different interpretation in different countries (UNESCO, 2011; de Vent et al., 2016b).

Third, there was some overlap in the studies that were used in Jewsbury et al. (2016) and the studies that were included in this factor meta-analysis, so the sample that was used to develop the model was not completely distinct from the sample used to evaluate its performance. Therefore, the two analyses were not independent, which could have artificially improved the performance of the CHC model.

To address these issues, in the next study, the factor models will be fitted to raw data from the Netherlands and Belgium, combined in the ANDI database. This database allows us to use a single test version for every variable, and to use a single standardized education scale. Also, because raw data are available, we can directly incorporate the influence of demographic variables on test variables, rather than using the more indirect approach of partialing out these variables from the correlations. Last, this is a completely different sample of studies from the samples used in study 1, and the samples used by Jewsbury et al. (2016).

5.4 study 2: factor analysis of the andi database

5.4.1 *Methods*

5.4.1.1 *Sample*

The construction and composition of the ANDI database are described elsewhere (de Vent et al., 2016a). This database includes data of studies that were conducted in the Netherlands and Belgium. For the data used in the present analysis, the number of included studies was 54, with a total of 11,881 participants. All test variables were transformed to normality in order to meet parametric assumptions and to speed up convergence, and were demographically corrected and standardized (de Vent al., 2016a). For the demographic corrections for level of education, we used a seven-point scale that is commonly used in Dutch neuropsychology (Verhage, 1964). This scale is comparable to the International Standard Classification of Education (UNESCO, 2011).

5.4.1.2 *Tests*

In study 2, the same test variables were included as in study 1. To remove the influence of test versions differing between studies, we included a single version for every test. Digit Span Forwards and Backwards were not included, as there were too few data for these variables for any specific version. LMI and LMII referred to Rivermead Behavioural Memory Test Stories Immediate Recall and Delayed Recall. SF referred to the Animals version of Semantic Fluency. COD referred to WAIS-III Digit Symbol-Coding. VLT referred to the Rey Auditory Verbal Learning Test.

5.4.1.3 *Model changes*

Because of the removal of the Digit Span subtests, the two versions of the CHC model collapse into a single version without "Working memory". The remaining factors were "Acquired knowledge or crystallized ability", "Processing speed", "Long-term memory encoding

Table 5.5: Factor Model Specifications of the Candidate Models for Study 2. Tests that Load on the Same Latent Factor Share a Letter. Some Tests Load on Multiple Latent Factors in the Hoogland and Jewsbury Models.

	TMTA	TMTB	LMI	LMII	LF	SF	COD	BNT	VLT-TR	VLT-DR
One factor	A	A	A	A	A	A	A	A	A	A
Strauss	D	D	C	C	B	B	A	E	C	C
Lezak	A	A	B	B	C	C	A	D	B	B
Gross	A	A	B	B	A	A	A	C	B	B
Hoogland	B	B + D	C	C	A + D	A + D	B	A	C	C
Larrabee	A	A	B	B	C	C	A	C	B	B
Jewsbury	B	B	A + C	A + C	E	E	B	A	C	C

Note. TMTA = Trail Making Test A, TMTB = Trail Making Test B, LMI = Logical Memory I, LMII = Logical Memory II, LF = Letter Fluency, SF = Semantic Fluency, COD = Digit Symbol Substitution / Coding, BNT = Boston Naming Test, VLT-TR = Verbal Learning Test - Total Recall, VLT-DR = Verbal Learning Test - Delayed Recall.

and retrieval", and "Word fluency". Like in study 1, factor loadings and covariances between latent variables were freely estimated. All latent variable variances were fixed to 1, so the covariances between latent variables can be interpreted as correlations. Residual variances of the tests are freely estimated as well.

The models were fitted using Mplus (Muthén & Muthén, 2012). Like in study 1, fit was evaluated by χ^2, RMSEA, SRMR, CFI, AIC, and BIC using the rules of thumb outlined in Schermelleh-Engel et al. (2003) to decide what constitutes bad, acceptable, and good fit.

5.4.2 *Results*

The Gross and Strauss models did not converge.. The Lezak model produced an error. The Jewsbury model converged, but produced a warning indicating a negative residual variance, which may indicate misspecification if the negative variance is large (Kolenikov & Bollen, 2012). However, the variance was not significantly different from 0, θ = -0.032, z = -0.581, p = 0.561.

The results of the model comparison between candidate models is given in Table 6. With respect to relative fit, the AIC and BIC indicate that the complex Jewsbury model fits better than the other models.

The χ^2 values indicates bad fit for all models (χ^2 / df > 3), except for the Jewsbury model, for which fit was acceptable (χ^2 / df > 2). All RMSEA values indicate good fit (all RMSEA < 0.05), except for the one factor model, for which the RMSEA indicates acceptable fit (RMSEA < 0.08). The SRMR values indicate bad fit for the one factor and Hoogland models (SRMR > 0.10), and acceptable fit for the

Table 5.6: Model Comparison Results.

	χ^2 (df)	RMSEA	SRMR	CFI	AIC	BIC
One factor	1647.0 (32)	0.065	0.149	0.736	73659.1	73880.6
Gross*	-	-	-	-	-	-
Hoogland	379.5 (24)	0.035	0.103	0.942	72407.6	72688.1
Lezak	368.3 (26)	0.033	0.092	0.944	72392.4	72658.2
Strauss*	-	-	-	-	-	-
Larrabee	370.5 (29)	0.031	0.095	0.944	72388.6	72632.3
Jewsbury	70.0 (24)	0.013	0.054	0.992	72098.1	72378.7

*Model did not converge or produced an error. - = not available from output.

Lezak, Larrabee, and Jewsbury models (SRMR > 0.05). The CFI values indicate bad fit for all models (CFI < 0.95), except for the Jewsbury model, for which fit was good (CFI > 0.97).

Next, we compared the CHC model fitted in study 2 to the CHC model fitted in study 1, to determine whether the factor structure was stable across the two analyses. The methods used in the two studies were dissimilar, i.e., correlation matrices served as the outcome measure in study 1 and actual test scores were the outcome measure in study 2. Because the scale of factor loadings and residual variances is dependent on the scale of the outcome measure, it is not warranted to compare factor loadings or residual variances between studies. However, the correlations between latent variables can be compared. To make the models comparable, the CHC model without the "Working Memory" latent variable from study 2 was fitted to the meta-analytic data from study 1 without DSF and DSB. The model is depicted in Figure 3, in which correlations between latent variables are also provided. Like in study 1, the "Processing Speed" factor is reverse coded. It can be seen that the correlations were in the same direction in both studies, and that correlations were lower for the second study. This could be due to the more appropriate demographic corrections: Regression-based corrections of the raw data were used rather than using a partial correlation approach, and level of education was coded on the same seven-point scale for all included samples.

5.5 general discussion

In this article, we sought to establish the cognitive domains that are measured by neuropsychological tests. Cognitive domains are used on a daily basis by neuropsychologists, to make decisions on which tests to administer to a particular patient, to determine whether a disorder affects a single domain or multiple domains, to calculate com-

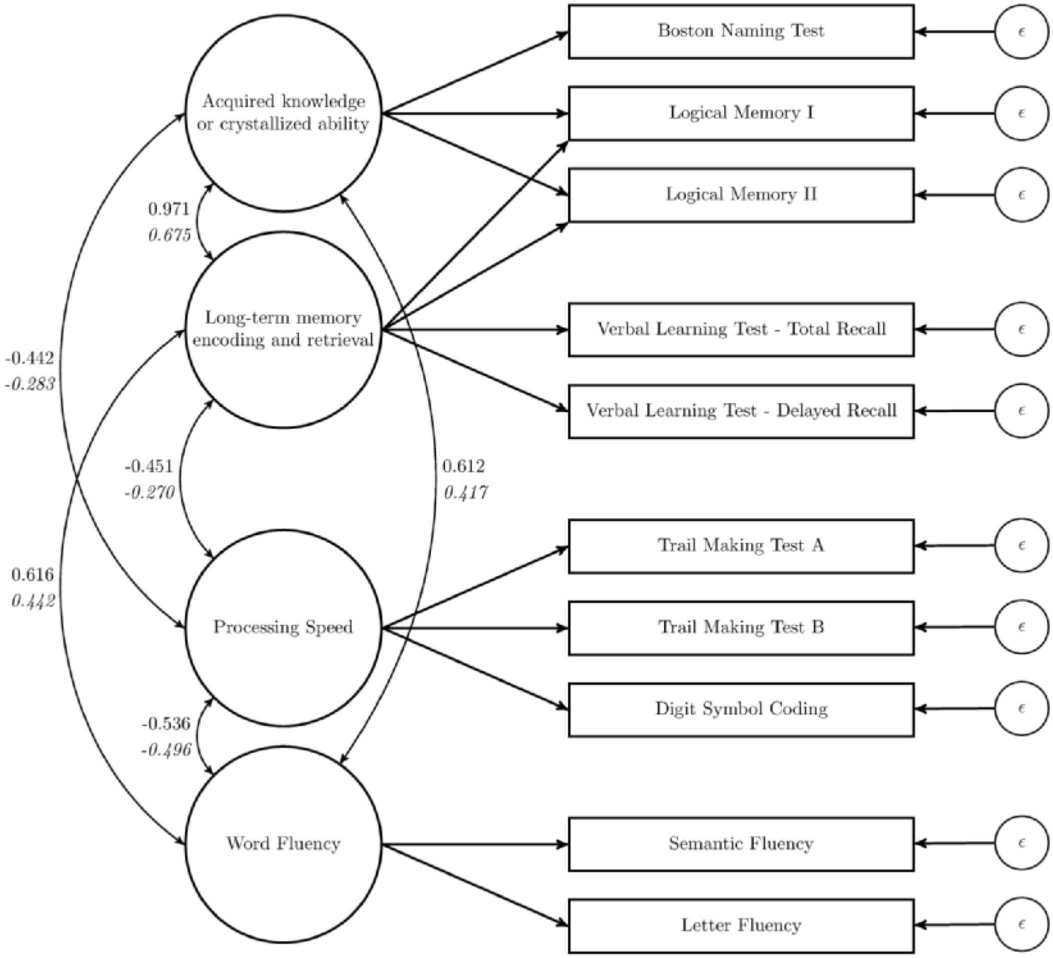

Figure 5.3: Jewsbury model for the ten tests included in study 2. For each combination of latent factors, the correlation is given for the meta-analytic data in roman type, and for the ANDI data in italic type.

posite scores of different tests belonging to the same domain, and to validate new tests that are designed to measure a particular cognitive function.

We compared several neuropsychological factor models that have been formulated in the literature. First, we performed a factor meta-analysis of correlation matrices, using the meta-analytic structural equation modeling framework (Cheung & Chan, 2005). Second, the different factor models were fitted to raw data from the ANDI database (de Vent et al., 2016a). Both analyses included a large number of neuropsychological tests, a very large sample, and accounted for the effects of age, sex, and level of education. Using these two different methods and samples, the same result was obtained: The Cattell-Horn-Carroll (CHC) model was shown to be the model that best described the data.

For the tests that were considered in this article, the CHC model consists of five intercorrelated factors: "Acquired knowledge or crystallized ability", "Long-term memory encoding and retrieval", "Processing speed", "Working memory", and "Word fluency". The Boston Naming Test and Logical Memory variables loaded on the first factor. The Verbal Learning Test variables and Logical Memory variables loaded on the second factor. Digit Symbol Substitution and Trail Making Test Parts A and B loaded on the third factor. The Digit Span variables and Trail Making Test Part B loaded on the fourth factor. Letter Fluency and Semantic Fluency loaded on the fifth factor.

The CHC model has three unique aspects compared to the other models fitted in this article. First, Letter Fluency and Semantic Fluency are typically paired with either the Boston Naming Test to form a "Language" factor (Larrabee) or are considered "Executive Functioning" tests (Strauss, Lezak, Gross, Hoogland). In the CHC model as formulated by Jewsbury et al. (2016), a separate factor is estimated for these fluency tests (Jewsbury & Bowden, 2016). Second, the Boston Naming Test is typically either a constituent of a "Verbal" factor (Larrabee, Hoogland) or is considered as separate from the other tests considered here (Strauss, Lezak, Gross). In the CHC model, the Boston Naming Test is paired with the Logical Memory variables to form the "Acquired knowledge or crystallized ability" factor. Third, the Digit Span variables are typically paired with Coding (Strauss, Lezak, Gross, Hoogland) and Trail Making Test Part A (Lezak, Gross, Hoogland). In the best-fitting CHC model, the Digit Span variables formed a separate factor and were not paired with any of these variables. Fourth, all other models, except for Hoogland, had no cross-loadings, i.e. all variables only belonged to one domain. The best-fitting CHC model had three cross-loadings, with the Trail Making Test Part B measuring both "Working memory" and "Processing speed", and Logical Memory Immediate Recall and Delayed Recall measuring both "Acquired knowledge or crystallized ability" and "Long-term memory encoding and retrieval".

Jewsbury et al. (2016) found that the CHC model provides a good fit for several datasets. The current study adds to the Jewsbury et al. findings in several ways. First, in two studies we were able to perform a single analysis of multiple datasets, thereby yielding a very large sample size. Second, the fit of the CHC model was good even though we corrected for age, sex, and level of education, which could have distorted earlier analyses. Third, we compared the CHC model to various alternatives, and even among those alternatives, the CHC model provided the best fit. Therefore, this article provides strong evidence for the CHC model.

The fact that the CHC model fits better than other models has a number of consequences for neuropsychology. First, a consequence of the cross-loadings in the CHC model is that it corroborates the view that tests generally measure more than one domain. For test selection, this does not mean that these are bad tests to administer, but rather that they can be informative for multiple domains at once. For example, if a low score on Trail Making Test Part B is observed, this could indicate impairment of "Processing speed" if observed with a low score on Trail Making Test Part A, and indicate impairment of "Working memory" if observed with a low score on Digit Span.

Second, the result has implications for the distinction between single-domain and multi-domain disorders. These disorders have typically been defined referring to the domains based on expert opinion, that is, "Executive Functions", "Memory", "Attention" etc. (Petersen, 2004). Given the results, it seems better to work instead with "Long-term memory encoding and retrieval", "Acquired knowledge or crystallized ability", "Processing speed", "Working memory", and "Word fluency". Application of the single-domain and multi-domain criteria with these domains would be straightforward. However, it is not clear whether the results that have been obtained in studies using the traditional domain definition (e.g., Ganguli et al., 2010; Libon et al., 2010) also hold with the CHC domain definition. It could be worthwhile to go back to already published data, and apply the criteria using the CHC domains to study their prognostic value in comparison to that of the criteria using the traditional domains. One important domain in terms of diagnosis in the traditional model is the "Memory" domain, which is used to define amnestic variants of disorders (Tabert et al., 2006). For the CHC model, the "Long-term memory encoding and retrieval" domain could be used for the same purpose, as all the same tests that load on the "Memory" factor load also on this factor.

Third, by calculating composite scores for a particular cognitive domain, one assumes that differences between people in their test scores are due to differences in their latent ability on this cognitive domain, i.e., that the cognitive domain is unidimensional (Borsboom, 2008). This is done for example in the calculation of an "Executive functioning" composite score (e.g. Gross et al., 2015), where one implicitly assumes that individual variation on Trail Making Test Part B, Coding, and Digit Span Backwards is due to individual variation in Executive functioning. We would advise against calculating such an "Executive functioning" composite score: The variables that are typically assigned to the "Executive Functioning" domain are spread out over three domains in the best-fitting CHC model ("Processing speed", "Working memory", and "Word fluency"), suggesting that unidimensionality is violated.

Fourth, it should be recognized that in both analyses, all latent factors were correlated in the CHC model. The influence of age and level of education that could have artificially produced such a correlation, had been partialed out. Therefore, although the tests in neuropsychological practice are designed to measure well-separable cognitive domains, these domains do not in fact seem completely separable. This could be due to the design of the tests. Perhaps tests have not been designed such that they can specifically measure individual variation only in "Working memory" while not also measuring variation in "Processing speed". However, this could also be due to the nature of cognitive functioning. All cognitive functions could be so deeply intertwined that it is not possible to measure one without the other (van der Maas et al., 2006).

It is important to realize the limitations of our results. First, the goal was to establish a factor model for cognitively healthy participants, but some participants included in the analyses may not have been cognitively healthy. Some of the contributing studies did not have the explicit goal to exclude pathology, but instead had the goal to obtain a representative sample from the population. This is true for both studies 1 and 2. Second, we should be careful not to overgeneralize the results to other samples. Tests loading on the same latent factor are not necessarily redundant measures of the same latent construct in all samples. For example, immediate recall and delayed recall on the Verbal Learning Tests were found to be indicators of the same latent factor in the CHC model. However, immediate and delayed recall are not interchangeable tests in clinical practice, as the function of one may be disrupted by disorder or injury while the other remains intact (Delis et al., 2003). Third, only part of the CHC factor model was tested in this study. Twelve variables were included in study 1 and ten variables were included in study 2, whereas many more test variables are used in clinical neuropsychology. With correlation matrices from newly published studies, the present meta-analysis could be extended to include other variables. To facilitate such an analysis, we provide correlation matrices in the appendix. We recommend that, as a rule, correlation matrices are shared publicly in articles or in supplemental materials, to facilitate the type of meta-analysis presented here.

To conclude, in two independent large-scale analyses the Cattell-Horn-Carroll (CHC) model best describes the structure of neuropsychological test domains. This model is more complex than models currently in use in neuropsychology, as it incorporates more domains, as tests load on multiple domains, and as domains are correlated. However, we have shown that such complexity is necessary to provide an accurate representation of cognitive functioning.

PREDICTING PARKINSON'S DISEASE DEMENTIA USING MODERN NEUROPSYCHOLOGICAL TECHNIQUES

6.1 ABSTRACT

Background: Parkinson's disease with mild cognitive impairment (PD-MCI) is a risk factor for the development of dementia (PDD) at a later stage of the disease. The consensus criteria of PD-MCI use a traditional test-by-test normative comparison. The aim of this study was to investigate whether a new multivariate statistical method allowing a formal evaluation of a patient's profile of test scores given a large aggregated database with regression-based norms, provides a more sensitive tool for predicting dementia status at three and five year follow up.

Methods: Cognitive test results of 123 newly diagnosed PD patients from a previously published longitudinal study were analysed with three different methods. First, the PD-MCI criteria were applied in the traditional way. Second, the PD-MCI criteria were applied using the large aggregated normative database. Last, multivariate normative comparisons were made using the same aggregated normative database. Progression to dementia after three and five years was used as a gold standard.

Results: The multivariate normative comparison was characterized by higher sensitivity and higher specificity in predicting progression to PDD at follow-up than the two PD-MCI criteria methods.

Conclusion: Modern statistical techniques allow for a more sensitive prediction of PDD than the traditional PD-MCI criteria.

6.2 INTRODUCTION

Many Parkinson's disease (PD) patients show a decline in cognitive functioning, often already early in the disease course (Aarsland et al., 2001; Hobson & Meara, 2004; Muslimovic et al., 2003). Mild Cognitive Impairment (PD-MCI) is predictive of further decline and progression to Parkinson's disease dementia (PDD; Aarsland et al., 2001; Caviness et al., 2007; Williams-Gray et al., 2007; Hoogland et al., 2017) It is important to accurately predict which patients will develop PDD as it

[0] Submitted as: Agelink van Rentergem, J. A.*, de Vent, N. R.*, Huizenga, H. M., Murre, J. M. J., ANDI Consortium, & Schmand, B. A. (2017). *Predicting Parkinson's disease dementia using modern neuropsychological techniques*. Manuscript submitted for publication.

may have implications for patient care, for example choice of medication (such as avoiding anticholinergic drugs) and planning of assistance. Also, accurate prediction enables a more appropriate selection of patients for cognitive interventions or pharmaceutical trials.

Clinical criteria for PD-MCI have been proposed by a task force of the International Parkinson and Movement Disorder Society (MDS; Litvan et al., 2012). In order to diagnose PD-MCI at level II (i.e. the level with most diagnostic certainty), a PD patient should experience subjective cognitive complaints (or their relatives should report such complaints) and should be impaired on objective cognitive testing. Litvan et al. (2012) recommend to administer at least two tests for each of five cognitive domains, thus a minimum of 10 tests, of which at least two tests need to indicate impairment before a PD-MCI diagnosis is set. Impairment is usually assessed by comparing the patient's test scores to those of normative samples, often in the form of norm tables that accompany published test manuals.

There are several issues with this way of working. First, each test has its own normative sample. Therefore, a patient is compared to different samples for each test. This means that the normative samples can differ in demographic composition, which means that a patient can be impaired on one test, and not another, merely because the samples against which the patient is compared are different. Second, since the normative data have been collected for each test separately, correlations between tests are usually unknown (except in rare cases of co-normed tests). Because the correlations are unknown, they cannot formally be taken into account in neuropsychological assessment. This makes it hard to evaluate abnormal combinations of scores (e.g. an abnormal score profile; Huizenga et al., 2007). Third, scores cannot always be corrected for the influence of demographic variables, even though age, sex and level of education are known to influence the scores on neuropsychological tests. It is often impossible to simultaneously correct for level of education, sex and age (Lezak et al., 2012). Also, when correction for age is possible, separate norms are presented for different age groups. When a patient gets older and shifts from one age group to the next, the interpretation of their test results can be different and may, for example, change from abnormal to normal (Zachary & Gorsuch, 1985). Fourth, when evaluating more than one test (at least 10 in the case of level II PD-MCI diagnosis), the likelihood of obtaining an abnormal score by chance alone increases with the number of tests that have been administered (Binder et al., 2009).

In this study, we applied a new statistical method to detect cognitive abnormality in newly diagnosed PD patients to predict PDD at later follow-up. This method uses an aggregated normative database of neuropsychological tests (de Vent et al., 2016).

Because the database contains data of co-normed neuropsychological tests, correlations between tests can be taken into account. This allows for a so-called multivariate normative comparison, which allows evaluation of a patient's profile of test scores. Multivariate normative comparison can detect abnormal combinations of high and low scores in a score profile, which are easily overlooked in traditional, univariate normative comparisons (Crawford & Garthwaite, 2002; Huizenga et al., 2007; Su et al., 2015). The database contains information on demographic variables and thus allows correcting for age, level of education and sex. By using regression-based demographic corrections, drastic changes in the interpretation of test scores when moving from one norm table to the next, are prevented. The new statistical method keeps the false positive rate under control, because it entails a single statistical comparison.

To examine whether this new approach is a good alternative to traditional (univariate) normative comparisons when predicting PDD, we compared its performance to that of the PD-MCI criteria. We used existing data from a longitudinal study conducted by our group (Broeders et al., 2013; Muslimovic et al., 2013; Broeders et al., 2013). First, we compared the ability of the traditional PD-MCI criteria to predict PDD after 3 and 5 years to the PD-MCI criteria when applied with a large normative database of co-normed tests. Second, we compared the traditional PD-MCI criteria to the new multivariate normative comparisons method when applied with the same large normative database. Finally, in supplement 2 we explored whether the new approach can give insight into which cognitive domains are impaired in PD-MCI patients who decline to PDD.

6.3 methods

6.3.1 *PD patients*

Participants were 123 patients with newly diagnosed PD (Muslimovic et al., 2005; Broeders et al., 2013) who at baseline were younger than 85 years, non-demented, had no history of stroke, and had a score of at least 24 on the Mini-Mental State Examination (MMSE; Lezak et al., 2012). Some patients did not participate in neuropsychological assessments after the baseline session but could still be included for the present analysis because information on their clinical status was available at the 3 and 5 years follow-up. After 3 years, the clinical status for 26 patients were missing. After 5 years, information was no longer available for another 24 patients. An overview of the demographic characteristics can be found in supplement 1.

6.3.2 *PD-MCI*

Broeders et al.15 applied the PD-MCI level II criteria (Litvan et al., 2012) as follows: 1) Patient has a PD diagnosis. 2) Patient, caregiver or clinician reports gradual cognitive decline. 3) Patient shows cognitive deficits on neuropsychological testing. 4) Cognitive deficits do not significantly interfere with functional independence. With respect to the first criterion, all patients in the sample were newly diagnosed PD patients; the diagnosis was checked by the study neurologists at follow-up. With respect to the second criterion, gradual cognitive decline reported by the patient was assessed by two questions, asking whether the patient experienced memory problems or concentration problems. If participants answered either question with "yes" or "sometimes", this was recorded as experiencing subjective complaints. With respect to the third criterion, a score of 1.5 SD below the demographically corrected mean on at least two tests was considered a cognitive deficit. With respect to the fourth criterion, patients were excluded if they had a score lower than 24 on the MMSE (Lezak et al., 2012). Finally, patients could also be diagnosed with PD-MCI if they reported no subjective complaints but had impairments (of 1.5 SD) at four or more tests.

6.3.3 *PDD*

PDD was used as the gold standard. PDD at 3 and 5 years follow-up was diagnosed by the MDS criteria.19 Criteria for PDD were defined as follows: 1) A diagnosis of PD prior to the onset of dementia. 2) an MMSE score lower than 24. 3) No depression. 4) Cognitive deficits severe enough to interfere with daily living, measured by the Barthel Activities of Daily Living (Collin, Wade, Davies, & Horne, 1988), Schwab & England Scale (Schwab & England, 1969), and Functional Independence Measure (Van Putten, Hornbart, Freeman, & Thompson, 1999). Also, an abnormal score on at least two of the following tests was required: clock drawing (Lezak et al., 2012), pentagon copying or serial 7s of the MMSE (Lezak et al., 2012).

6.3.4 *Materials*

PD patients were tested on five cognitive domains: memory, language, executive functions, visuospatial skills and attention. All test variables from the Broeders et al. (Broeders et al., 2013) study were used except the Modified Wisconsin Card Sorting Test (Lezak et al., 2012) as its score distribution was extremely skewed, violating the assumptions of the parametric normative comparisons that are used throughout this article. We substituted it by the Tower of London (Lezak et

Table 6.1: Characteristics of the Neuropsychological Test Variables in ANDI.

	N	% Men	Age range	Demographic variables[a]
Memory				
Rey Auditory Verbal Learning Test - total	5017	50	18-97	S + A + E
Rey Auditory Verbal Learning Test - delayed recall	4540	49	18-97	S + A + E
Rivermead Behavioral Memory Test - Story subtest - immediate recall	346	40	19-90	S + A + E
Rivermead Behavioral Memory Test - Story subtest - delayed recall	353	40	19-89	S + A + E
Language				
30-item Boston Naming Test	467	42	18-89	S + A + E
WAIS-III Similarities	274	36	18-80	E
Executive functions				
Controlled Oral Word Association Test	2894	48	18-97	S + A + E
Tower of London - total movement score	62	53	40-80	A
Visuospatial/constructive skills				
Judgement of Line Orientation	69	54	40-80	S + E
Clock Drawing Test	167	46	40-82	E
Attention				
WAIS-R Digit Symbol Test	2122	43	18-91	S + A + E
Trail Making Test - part A	3216	47	18-97	S + A + E

[a] As explained elsewhere (de Vent et al., 2016), an AIC selection procedure was used to estimate which of the three demographic variables to include in regression-based demographic corrections. In this column, S, A and E indicate whether sex, age and level of education were included for each variable.

al., 2012) as an alternative test for the executive functions domain. An overview of the tests can be found in Table 1.

6.3.5 *Normative control sample*

For normative comparisons we used either the published norms of each neuropsychological test or the database of the Advanced Neuropsychological Diagnostics Infrastructure (ANDI; de Vent et al., 2016). ANDI is an online tool that can be used by clinicians and researchers to conduct normative comparisons. ANDI has a large aggregated normative database (N=26,939) which consists of participants that either participated as healthy control subjects in clinical studies, or participated in community-based studies. Since each participant completed only a subset of the tests which are included in ANDI, the number of participants per test varies between 69 and 5783 depending on the test. Table 1 provides the number of participants per test and demographic information. For most test variables, sex, age and level of education had a significant effect, and were included in the demographic correction (de Vent et al., 2016).

6.3.6 *PD-MCI criteria applied with ANDI's normative data*

In applying the PD-MCI criteria, Broeders et al. (2013) followed typical neuropsychological practice and used normative data from test manuals and various other sources to judge whether a patient deviated from the norm. Here, we applied the PD-MCI level II criteria in the same way but now with the ANDI database instead of the normative data accompanying each test. A difference between the norms is that the ANDI data have been treated in a consistent manner across all tests (de Vent et al., 2016). This includes uniform procedures of outlier removal, test score standardization, and selection of transformations to normality. Also, for many tests, a larger normative sample is available. Student's t-statistics were used in calculating whether scores were abnormal (Crawford & Garthwaite, 2002). A threshold p-value of 0.067 one-tailed was used to define impairment, which corresponds to -1.5 SD below the mean. Because tests were one-tailed, only deviations in the negative direction were classified as impaired.

6.3.7 *Abnormality as defined by MNC*

Finally, we examined the performance of multivariate normative comparisons (MNC) 8. MNC compares the profile of the patient's scores to the norm, i.e., to the profile of scores that is predicted for a healthy participant of the same sex, age and level of education (Agelink van Rentergem et al., 2017; Agelink van Rentergem et al., 2017). MNC result in a p-value, which indicates abnormality when it is below a certain threshold. We tested for impairment (one-sided), i.e., only deviations in the negative direction were classified as impaired. In univariate comparisons, if the patient had no subjective complaints, we required four instead of two significant deviations. In MNC this adaptation is not possible, as only a single comparison is performed. Therefore, we used different threshold values for those with and without subjective complaints. For patients without subjective complaints, a threshold p-value of 0.067 was used. For patients with subjective complaints, a more lenient threshold p-value of 0.134 was used.

6.3.8 *Analysis*

We calculated whether the classification at baseline is predictive of developing PDD. Sensitivity and specificity were compared across the three methods: PD-MCI criteria, PD-MCI criteria with ANDI, and MNC applied with ANDI. Sensitivity was calculated by dividing the

Table 6.2: Demographic and Clinical Characteristics for the Three Groups (PD-MCI criteria, ANDI PD-MCI criteria and ANDI MNC) at Baseline.

	PD-MCI criteria	ANDI PD-MCI criteria	ANDI-MNC
	normal cognition	PD-MCI	normal cognition
	N = 80 (65%)	N = 43 (35%)	N = 90 (73%)
Age	65.1 (10.6)	68.0 (10.1)	64.4 (10.7)
Sex M/F	43/37	23/20	46/44
MMSE	28 (1.9)	27 (2.0)	28.1 (1.9)
Disease duration in months	18.3 (8.9)	20.1 (13.4)	18.0 (8.8)
LED	139.0 (142.6)	149.9 (139.3)	139.1 (143.5)
UPDRS	15.8 (7.8)	19.4 (7.8)	16.2 (8.2)
H&Y	1.6 (0.7)	2.1 (0.7)	1.7 (0.7)
HADS	8.5 (6.6)	13.5 (7.5)	9.4 (7.1)
SE-ADL	91.2 (5.8)	88.1 (7.9)	90.6 (6.9)
BADL	19.7 (0.7)	19.4 (1.5)	19.6 (1.2)

Abbreviations: PD-MCI = Parkinson's Disease Mild Cognitive Impairment; MNC = Multivariate Normative Comparisons; MMSE = Mini-Mental State Examination; LED = Levodopa Equivalent Dose; UPDRS = Unified Parkinson's Disease Rating Scale; H&Y = Hoehn & Yahr scale; HADS = Hospital Anxiety and Depression Scale; SE-ADL = Schwab & England Activities of Daily Living; BADL = Behavioral Assessment of Daily Living.

number of patients who were classified as impaired at baseline and develop PDD, by the total number of patients who developed PDD. Specificity was calculated by dividing the number of patients who were classified as not-impaired at baseline and did not develop PDD, by the total number of patients who did not develop PDD. This was done separately for the development of PDD after three years, and after five years.

6.4 RESULTS

6.4.1 *Demographic characteristics*

In Table 2, demographic and clinical characteristics are given for the patients, separated into cognitively normal and abnormal categories using each of the three methods.

6.4.2 *Progression to PDD*

Figure 1 shows the progression to PDD for each method. With the criteria used by Broeders et al. (2013) , 35% of the PD patients had

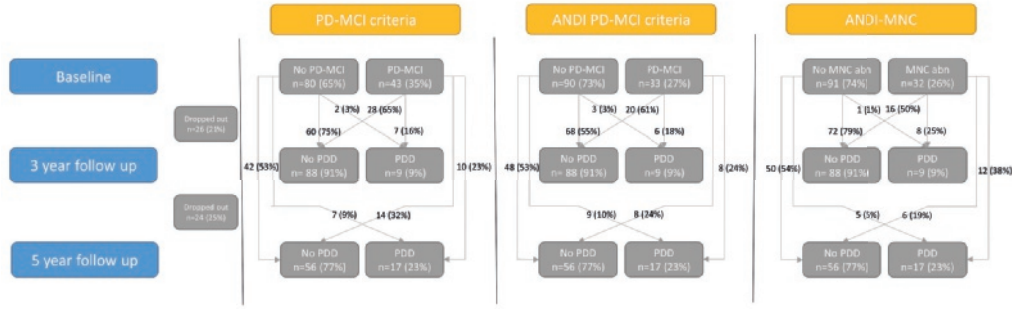

Figure 6.1: Progress of PD patients (n = 123) to PDD after 3 (n = 97) and 5 years (n = 73) for the three methods; PD-MCI criteria, PD-MCI criteria applied with ANDI, and multivariate normative comparisons (MNC) applied with ANDI.

PD-MCI at baseline. After three years, 16% of the PD-MCI patients had progressed to PDD and 65% had not. Of the group who did not have PD-MCI, 3% of patients nevertheless had progressed to PDD and 75% had not (the remaining patients were lost to follow-up; see supplementary materials). After five years, 23% of those with PD-MCI at baseline had progressed to PDD while 32% had not. Of the group who did not have PD-MCI, 9% had progressed to PDD while 53% had not.

The PD-MCI criteria applied with ANDI show that 27% of the patients had PD-MCI at baseline. After three years, 18% of the PD-MCI patients had progressed to PDD and 61% had not. Of the group who did not have PD-MCI, 3% had progressed to PDD and 55% had not. After 5 years, 24% of the PD-MCI patients had progressed to PDD and 24% had not. Of the group who did not have PD-MCI, 10% patients had progressed to PDD while 53% had not.

The multivariate normative comparisons (MNC) method applied with the ANDI normative data shows that 26% PD patients were considered to be MNC-impaired at baseline. After 3 years, 25% of the MNC-impaired PD patients had progressed to PDD and 50% had not. Of the group who were not MNC-impaired, 1% had progressed to PDD and 79% had not. After 5 years, 38% of the MNC-impaired PD patients had progressed to PDD and 19% had not. Of the group who were not MNC-impaired, 5% patients nevertheless had progressed to PDD while 54% had not.

In Figure 1, it is not visible how much overlap there is in the three different types of diagnostic methods at baseline. For example, the 32 classified as MNC-impaired could be different patients from those 33 classified as having PD-MCI using ANDI. The overlap in diagnoses between pairs of classification methods is explored in supplement 3. Each of the three methods did indeed differ somewhat in the patients

Table 6.3: Sensitivity and specificity for PDD of each method (original PD-MCI criteria, PD-MCI applied with ANDI, and MNC method applied with ANDI), specified for 3 and 5 year follow up. In parentheses: 90% confidence interval.

	3 year follow-up		5 year follow-up	
	sensitivity	specificity	sensitivity	specificity
PD-MCI criteria	0.78 (0.50-0.93)	0.68 (0.60-0.76)	0.59 (0.39-0.76)	0.75 (0.64-0.83)
PD-MCI criteria ANDI	0.67 (0.40-0.86)	0.77 (0.69-0.84)	0.47 (0.29-0.66)	0.86 (0.76-0.92)
MNC ANDI	0.89 (0.61-0.99)	0.82 (0.74-0.88)	0.71 (0.50-0.85)	0.89 (0.80-0.95)

they classified as impaired, although the percentages of agreement were high (78-87%) and kappa's ranged from 0.49 to 0.68.

6.4.3 *Sensitivity and Specificity*

Sensitivity and specificity of the three methods are given in Table 3.

6.5 discussion

We investigated three methods for detecting cognitive abnormalities in PD-patients that predict progression to PDD. We compared the predictive performance of the PD-MCI criteria, applied either with traditional normative data (Broeders et al., 2013) or with the ANDI normative database, to the performance of MNC using the ANDI database. We found that the number of patients diagnosed with PD-MCI at baseline differed between these methods. The original PD-MCI criteria as applied by Broeders et al. 15 resulted in 35% of the PD patients being diagnosed with PD-MCI. Using the same criteria but with ANDI normative data, this decreased to 27%. The MNC method applied with ANDI concluded that 26% of the patients were cognitively abnormal at baseline. In the literature, the frequency with which cognitive impairments in PD patients are reported differs greatly between studies (probably due to differences in methodology and in sample characteristics, such as disease duration or severity). Studies with comparable methods (1.5 SD deviations on at least two out of ten tests) show that between 21% and 60.5% of PD patients are diagnosed with PD-MCI (Janvin, Aarsland, Larsen, & Hugdahl, 2003; Hobson & Meara, 2015; Gasca-Salas et al., 2014; Domellöf, Ekman, Forsgren, & Elgh, 2015; Santangelo et al., 2015; Galtier, Nieto, Lorenzo, & Barroso, 2016; Pedersen, Larsen, Tysnes, & Alves, 2017). The new multivariate normative comparison technique yields a number that lies at the low end of this range.

In terms of prediction, the MNC method applied with the ANDI database performed best. Sensitivity and specificity were higher for this method than for the two PD-MCI criteria methods. This was true for both the prediction of PDD after three and after five years. The MNC method applied with ANDI leads to a more sensitive detection of cognitive pathology that cannot readily be obtained with conventional diagnostic methods, which seems mainly to be due to use of a multivariate statistical technique and not to use of a large aggregated database. Between the two PD-MCI criteria methods, there was little difference in terms of accuracy. The PD-MCI criteria applied with ANDI resulted in a slightly lower sensitivity and a slightly higher specificity compared to the PD-MCI criteria as applied by Broeders et al. (2013). Just using the ANDI database instead of traditional norms therefore does not seem to improve prediction by itself.

Figure 2 gives an overview of the sensitivity and specificity found in previous studies that also used 1.5 SD as a cutoff score. Previous studies reported a sensitivity of the PD-MCI criteria for PDD ranging from .52 (Pedersen et al., 2017) to .92 (Gasca-Salas et al., 2014) and specificity ranging from .46 (Galtier et al., 2016) to .94 (Hobson & Meara, 2015). Therefore, the sensitivity and specificity estimates obtained with the MNC are at the high end of the spectrum.

In Figure 2, for all three methods a decrease in sensitivity can be observed between the 3 year and 5 year follow-up. An explanation would be that with a short period between baseline and PDD diagnosis, most patients who developed dementia were already impaired, leading to a high sensitivity. With more time between baseline and PDD diagnosis, some patients who developed dementia may have been unimpaired at baseline, leading to a lower sensitivity. Similarly, a small increase in specificity between the 3 year and 5 year follow-up can be observed. This is explained by the time it takes to develop dementia: Patients who are impaired at baseline may still not progress to dementia in the first few years after baseline, leading to a low specificity. As more time passes however, patients who were impaired at baseline will probably develop dementia, leading to an increase in specificity.

There are several limitations to our study. The number of patients was not very large (n=123) and loss to follow-up was quite high (21% at 3 years, and another 25% at 5 years). However, the numbers lost to follow-up are not different between those cognitively normal or abnormal at baseline (in supplement 3 an specification of which patients were lost to follow up is given).

Subjective complaints were used in PD-MCI criteria and MNC. Therefore, subjective complaints played a large role in determining the diagnoses in this study, while they were established using only two questions. Possibly, higher specificity and sensitivity would have been obtained, had we established subjective complaints more formally,

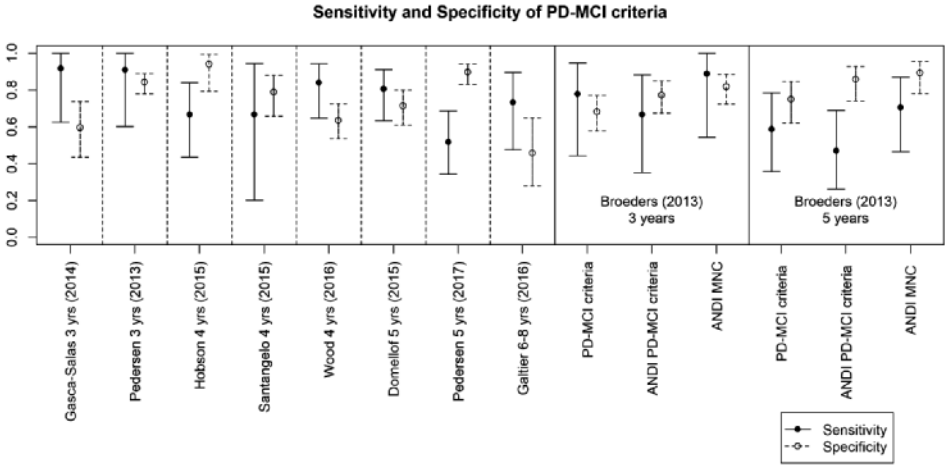

Figure 6.2: Sensitivity and specificity of the PD-MCI criteria for PDD when using 1.5 SD as a cutoff score in various previous studies (left panels), and sensitivity and specificity of the three methods investigated in the current study (right panels). Error bars indicate 95% confidence intervals 43.

for example with a longer, validated questionnaire, ideally including reports by relatives, caregivers and clinicians. Instead, for patients without subjective complaints, deviation on at least four neuropsychological tests was used as a criterion for PD-MCI, and a more strict criterion was used for MNC.

In sum, we conclude that the multivariate normative comparison method enables a better prediction of who will progress to dementia than the conventional PD-MCI method.

7

UNIVARIATE COMPARISONS GIVEN AGGREGATED NORMATIVE DATA

7.1 ABSTRACT

Objective: Normative comparison is a method to compare an individual to a norm group. It is commonly used in neuropsychological assessment to determine if a patient's cognitive capacities deviate from those of a healthy population. Neuropsychological assessment often involves multiple testing, which might increase the familywise error rate (FWER). Recently, several correction methods have been proposed to reduce the FWER. However these methods require that multivariate normative data are available, which is often not the case. We propose to obtain these data by merging the control group data of existing studies into an aggregated database. In this paper we study how the correction methods fare given such an aggregated normative database.

Methods: In a simulation study mimicking the aggregated database situation, we compared applying no correction, the Bonferroni correction, a maximum distribution approach and a stepwise approach on their FWER and their power to detect genuine deviations.

Results: If the aggregated database contained data on all neuropsychological tests, the stepwise approach outperformed the other methods with respect to the FWER and power. However, if data were missing, the Bonferroni correction produced the lowest FWER.

Discussion: Overall, the stepwise approach appears to be the most suitable normative comparison method for use in neuropsychological assessment. When the norm data contained large amounts of missing data, the Bonferroni correction proved best. Advice of which method to use in different situations is provided.

7.2 INTRODUCTION

Normative comparison is a method of comparing test scores of an individual to those of a norm group. It is often applied in neuropsychological assessment, with the goal to draw conclusions about an individual's cognitive capacities, like memory or attention. If an individual deviates sufficiently from the norm group, a group of healthy individuals, we may speak of 'abnormality' (Crawford & Howell,

[0] Published as: Zadelaar, J. N.*, Agelink van Rentergem, J. A.*, & Huizenga, H. M. (2017). Univariate comparisons given aggregated normative data. *The Clinical Neuropsychologist*, *31*(6-7), 1155-1172.

1998; Harvey, 2012; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999; Lezak et al., 2012). As such conclusions may affect one's academic, professional and personal life, assessment accuracy is vitally important. For example, a 'healthy' individual being falsely diagnosed with cognitive impairments could result in a waste of time and treatment resources, as well as personal suffering. Similarly, an undiagnosed condition may linger or worsen over time, possibly with dire consequences for the individual and her/his surroundings (Harvey, 2012). As such, the focus of this paper will be on improving statistical methods for normative comparison, as used in neuropsychological assessment.

In neuropsychological assessment it is common to administer multiple tests (Harvey, 2012). However, multiple testing is associated with an increased chance of at least one test falsely indicating abnormality, that is, with an increased familywise error rate (FWER) (Binder, 2009; Feise, 2002; Huizenga et al., 2016; Huizenga et al., 2007; Van der Laan, Dudoit, & Pollard, 2004). In terms of neuropsychological assessment, this means that administering more tests to an individual increases the chance of at least one test falsely indicating cognitive abnormality. Therefore, methods that correct for an increased FWER should be applied.

Unfortunately, FWER corrections may decrease the ability to detect true deviations (Verhoeven, Simonsen, & McIntyre, 2005). In neuropsychological assessment, this means that a method's ability to detect real cognitive abnormalities decreases. Still, both a low FWER and a high power to detect true deviations are important for good assessment accuracy. As such, the goal of this study is to develop a normative comparison method that successfully reduces the increased FWER associated with multiple testing while not sacrificing too much power. Three candidate methods will be examined: the well-known Bonferroni correction, and two new methods, the maximum distribution approach, and the stepwise approach.

The Bonferroni correction reduces the increased FWER caused by multiple testing. This method is often favored for its simplicity (Armstrong, 2014; Cao & Zhang, 2014) but is also known for its excessively low power when tests are correlated (Bland & Altman, 1995; Moran, 2003; Narum, 2006; Verhoeven et al., 2005). As such, this method is not expected to perform well, but is included nonetheless due to its simplistic nature.

Next is the maximum distribution approach (or max-approach, for short), which also reduces FWER. (Huizenga et al., 2016; Nichols & Holmes, 2002). An advantage this method has over the Bonferroni correction is that it better retains power when tests are correlated (Huizenga et al., 2016). This is expected to improve assessment accuracy.

The stepwise approach also reduces FWER, and increases power even further (Huizenga et al., 2016; Nichols & Holmes, 2002). Notably, this method is the most demanding computationally. However, it can be implemented in user-friendly software.

One problem all these methods face though is requiring an appropriate norm group. After all, comparing an 80-year old male to a norm group of 20-year old females may well result in deviation(s) attributable to demographic differences rather than cognitive abnormalities. As such, neuropsychological assessment requires a norm group that either: 1) consists solely of people from a similar demographic background as the assessed individual, or 2) is sufficiently large and varied to correct for such influences (Crawford & Howell, 1998). Additionally, the max approach and stepwise approach require that multiple participants in the normative sample performed on all tests that were administered to the individual (Huizenga et al., 2016). Such a normative sample will rarely be available. In order to provide a solution, we propose to merge already available datasets – the 'healthy' control groups of previously conducted studies – to create one dataset that meets these demands (de Vent et al., 2016; Agelink van Rentergem et al., 2017; Agelink van Rentergem et al., 2017). With data-sharing increasing in popularity in the social sciences (Asendorpf et al., 2013; King, 2011; Poline et al., 2012; Vines et al., 2014), this seems like an opportune solution to the appropriate norm group problem.

Aggregating studies like this results in a multilevel dataset with two levels; a participant and a study level, with the former nested within the latter (Steenbergen & Jones, 2002). This creates two potential problems. First, the dataset now contains both within-study variance and between study-variance, as opposed to only within-study variance. If and how this might affect the assessment accuracy (i.e. the FWER and power) of the aforementioned methods is yet unclear. Second, not every included study contains every test of interest, resulting in systematically missing data, which may also affect assessment accuracy (Dupont & Plummer, 1990; Field, 2009). This is why the accuracy of normative comparison methods when applied to multilevel structured data with missing data needs to be examined.

Huizenga et al. (2016) investigated whether Bonferroni correction, max-approach and stepwise approach normative comparison methods based on resampling adequately corrected for multiple testing if the normative database was of a non-aggregated nature. In the current study, we adapted Huizenga et al.'s max-approach and stepwise approach to the aggregated database case by including empirical instead of resampled distributions. Both are non-parametric methods, and therefore require fewer assumptions than those based on theoretical distributions (Nichols & Holmes, 2002). This imposes less restrictions on the norm dataset, making the methods more flexible in application. A difference is that the resampling methods of Huizenga et al. (2016) perform well with small samples sizes, whereas the current methods based on empirical distributions require a norm database consisting of many participants, which fortunately is the case in the suggested aggregated database case. An advantage of the current methods is that they: 1) can easily be extended to aggregated data as described above and 2) that they are computationally and theoretically simpler than the resampling methods, making them more user-friendly and easily interpretable.

Uncorrected normative comparison, and normative comparison with the Bonferroni correction, the max-approach and the stepwise approach were applied to non-multilevel and multilevel data, with and without missing data, while varying a number of data parameters, such as the number of tests and norm group sample size. Accuracy was estimated by calculating the FWER and power. The uncorrected method was expected to produce an increased FWER whenever multiple testing occured. All FWER correction methods were expected to produce FWERs that: 1) were lower than the FWERs of the uncorrected method, and 2) approximated the preset significance threshold ($\alpha = 0.05$). Amongst the correction methods, the stepwise approach was expected to produce the highest power. The Bonferroni correction was expected to produce the lowest power when tests were correlated. The power of the new correction methods was aimed to equal or exceed that of the Bonferroni correction.

7.3 METHODS

7.3.1 *Normative Comparison Methods*

This section explains the aforementioned methods for normative comparison on a more detailed level. Normative comparison entails comparing a single test score to the distribution of a norm group's test scores. In the uncorrected normative comparison, this requires calculating the proportion of norm group scores on a certain test that are more extreme than the individual's score on this same test; this proportion constitutes the p-value of the individual's test score . If this p-value falls below the preset significance threshold ($p < \alpha$), we may conclude that the individual deviates significantly from the norm group on the tested cognitive capacity. This is done separately for each of the M administered tests; when $M = 1$, the FWER equals the threshold, FWER = α; if $M > 1$, the FWER increases, FWER > α (Feise, 2002; Huizenga et al., 2016).

To counter the increased FWER caused by multiple testing, normative comparison can be augmented with the Bonferroni correction. This correction entails implementing a new, stricter significance threshold, which is calculated by dividing the original threshold by the number of performed tests: α_Bonferroni= α/M. This results in a more stringent significance threshold as the number of tests increases. With a more stringent threshold, more extreme scores are required to produce a significant result, thus reducing the FWER. This correction is computationally easy and performs well when tests are not correlated amongst each other. Unfortunately, when tests are correlated it becomes too conservative, as the Bonferroni correction corrects as if the tests were uncorrelated, resulting in overcorrection (Bland & Altman, 1995; Holm, 1979; Narum, 2006). This causes an unnecessarily large decrease in both FWER and power, with the latter posing a problem for this method's accuracy.

Unlike the Bonferroni correction, the max-approach does not correct the significance threshold but instead changes the norm group distribution. That is, an individual's test scores are not compared to the distribution of the norm group scores on the corresponding test – as is done in uncorrected normative comparison – but instead to the max-distribution. This max-distribution is obtained by taking every norm group participant's most extreme score over all M tests, and combining these scores into one distribution. As a result, the max-distribution contains only the most extreme norm group scores. If an individual's scores deviate significantly even when compared to these most extreme scores of a norm group, it is more likely to reflect true deviation. As such, the max-approach reduces FWER (Blakesley et al., 2009; Huizenga et al., 2016; Nichols & Holmes, 2002; Westfall & Young, 1993). An advantage this method has over the Bonferroni correction is that it takes into account test correlations. This prevents the overcorrection associated with correlated tests, allowing for FWER correction while not sacrificing too much power, resulting in better accuracy.

The stepwise approach starts by ordering the individual's M test scores and comparing the most extreme score to the max-distribution. All other scores are compared to the max-distribution over all tests, not including the ones corresponding to more extreme scores. That is, the second most extreme score is compared to the max-distribution over all tests except the one corresponding to the most extreme score, the third-most extreme score is compared to the max-distribution over all tests except the tests corresponding to the most and second-most extreme scores, etc. Like the max-approach, the stepwise approach reduces FWER by requiring more extreme results to obtain significance, while maintaining power by taking into account between-test correlations. Unlike the max-approach though, it compares less extreme scores to less extreme distributions, meaning these scores have a higher chance of reaching significance. This increases the power even further (Gordon & Salzman, 2008; Huizenga et al., 2016; Westfall & Young, 1993) .

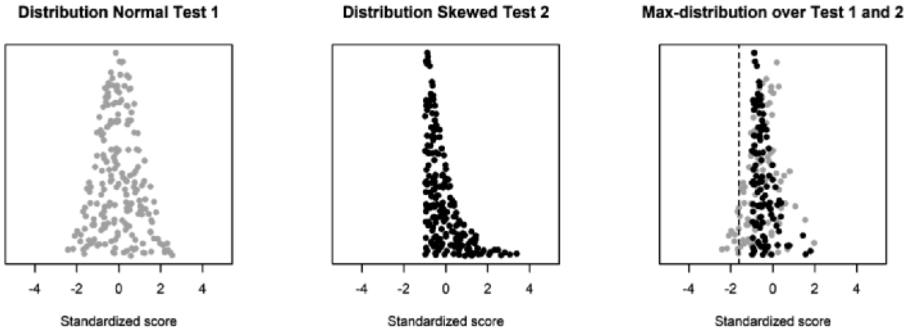

Figure 7.1: Example of how the max-distribution is affected by tests having different distributions; when one test is normally distributed (left), the other is skewed to the right (middle). The dotted line indicates the critical value at α=0.05.

Both the max-approach and stepwise approach require standardized scores, as using unstandardized scores causes tests with a more extreme scoring range (e.g. the number of seconds required in a Stroop task) to dominate the max-distribution, disallowing tests with a smaller scoring range (e.g. the number of errors in a Stroop task) from becoming significant.

Additionally, the max-approach and stepwise approach require norm group scores to be similarly distributed across tests. If not, tests with skewed distributions may be over- or underrepresented. Figure 1 shows a test with a normal distribution, a test with a skewed distribution, and the max-distribution the pair of tests produce. Herein, only scores from the normally distributed test are represented in the lower tail, beyond the critical value. As such, on the second (skewed) test, the assessed individual requires a score excessively extreme compared to the corresponding test's norm distribution to be found significant, thus lowering the power. Should norm group test score distributions be found to substantially differ, transforming the data to normality is recommended (de Vent et al., 2016).

In the following paragraph, we outline how we compared these methods in a simulation study.

7.3.2 *Data Simulation*

Data were simulated in *R* (R Core Team, 2015), with each dataset containing normative data (the norm group) and patient data (the assessed individual). Normative data were simulated as if the data from one or more studies (non-multilevel vs. multilevel data), each containing some or all of the possible tests (no missing data vs. missing data), were merged. In creating the datasets, the following parameters were varied: the number of studies (S), the number of participants per study (N), the number of tests (M), the between-test correlations (BTC), the between-study variance (BSV), and the number of tests in the patient data that showed deviation. Parameter settings were based on the Advanced Neuropsychological Diagnostics Infrastructure (ANDI), a recent initiative in neuropsychological diagnostics containing healthy participant data of various neuropsychological tests, as collected from multiple studies (de Vent et al., 2016; http://www.andi.nl/home).

7.3.2.1 *Number of Tests (M): {1, 2, 3, 5, 15, 24, 50}*

The number of administered tests was based on the mean number of tests per study in ANDI, resulting in M = 15; M = 24 was chosen to represent a larger, yet still realistic number of tests. We chose M = 50, as to investigate the effect an extremely large – albeit unrealistic – number of tests had on the analyses. Similarly, M = 2, M = 3 and M = 5 were chosen to investigate hypothetical situations with a relatively small number of tests. Finally, M = 1 served as a baseline, illustrating each method's performance when multiple testing did not occur.

7.3.2.2 *Number of Studies (S): {1, 2, 5, 20, 40}*

The mean number of studies in ANDI to include at least one common test was 18, and the largest number was 37. Rounding upwards this became S = 20 and S = 40; S = 1 was included to investigate how each method performed when applied to non-multilevel data; S = 2 and S = 5 were added to examine the effect of multilevel data made up of a small number of studies.

7.3.2.3 *Number of Participants per Study (N): {10, 20, 70, 200}*

The number of participants greatly varies within the ANDI database, as data sources vary from large community samples, to small matched samples in studies about rare diseases. We based our typical sample size on the latter, and chose N = 70. The minimum and maximum number of participants per study of N = 10 and N = 200 were based on the smallest and largest number of participants per study observed in the ANDI data, omitting the large community samples. Data were simulated as if all studies had the same number of participants.

7.3.2.4 *Between-Test Correlations (BTC): {0, 0.27, 0.5, 0.8}*

Between-test correlations describe the correlations between tests from the same study; BTC = 0.27 was the mean between-test correlation in ANDI, and BTC = 0.8 was the largest between-test correlation. Given the large difference between these values – mostly attributable to the unusually large value of 0.8 – BTC = 0.5 was added as to illustrate the effect of high but still common between-test correlations. Additionally, BTC = 0 was chosen to include a situation with completely uncorrelated tests.

7.3.2.5 *Between-Study Variances (BSV):{0, 0.15, 0.4}*

Between-study variance describes the variance in the norm group dataset attributable to differences between studies, leaving remaining variation attributable to individual differences. These values were based on the intra-class correlations (ICC) found in ANDI; 0.15 was the mean ICC in ANDI, and 0.4 the largest ICC found in this dataset. From these correlations, the between-study variances could be computed through the formula: BSV=ICC*σ^2, wherein σ^2 equals the total variance of the norm group dataset (Tabachnick & Fidell, 2007). In each dataset, σ^2 was arbitrarily set to 1, resulting in BSV = 0.15 and BSV = 0.4. BSV = 0 was included to examine a situation wherein all studies involved were completely equivalent.

7.3.2.6 *Missing Data: {0%, 50%}*

The amount of missing data was set at either 0% (no missing data) or 50% (half of the data were missing). The latter was deemed a sufficiently large percentage to demonstrate the effects of missing data, and was computed by removing scores after data simulation. This was done by removing the first half of the tests (test 1 to M/2) for the first half of the studies (study 1 to S/2), and removing the second half of the tests (test M/2 + 1 to M) from the second half of the studies (study S/2 + 1 to S), as illustrated in Figure 2.

7.3.2.7 *Patient Deviation: {1; 5}*

The number of tests a patient could deviate on was varied to illustrate the expected increase in power of the stepwise approach over the max-approach in situations with multiple deviating tests. The patient could deviate on either the first, or on the first five tests.

7.3.2.8 *Norm Data Simulation*

The norm group data were simulated as if test scores had already been corrected for demographic influences, meaning they had a mean of zero (de Vent et al., 2016). Thus, the scores of the norm group data only consisted of a within-study term epsilon (ϵ) and a between-study term denoted by nu (ν). Epsilons differed for each participant and each test. Nu's differed for each study and each test. By adding these two elements, the test scores were computed: score = $\epsilon + \nu$. Epsilons were drawn from a multivariate normal distribution with means of zero and a covariance matrix with variances of 1-BSV and covariances calculated with the BTC values. Nu's were drawn from a

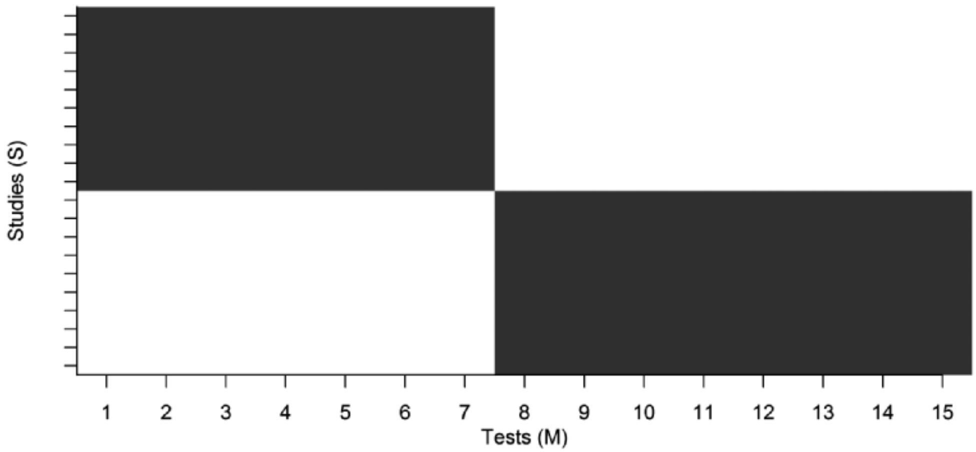

Figure 7.2: Missing data pattern with 50% of the data missing. Grey areas indicate non-missing values, white areas indicate missing values.

multivariate normal distribution with means of zero and a covariance matrix with variances of BSV and covariances of 0. Note that because a non-multilevel dataset consists of only one study (S = 1), it should have no between-study variance (BSV = 0), causing the nu's to equal zero, meaning non-multilevel scores consisted solely of epsilons.

7.3.2.9 *Patient Data Simulation*

Patient data had the same format as the norm dataset, but for N = 1. Patients were either healthy (with scores equaling the mean used in simulating the norm data) or deviant (two standard deviations below the mean used to simulate the normative data, either on the first test or on the first five tests). The inclusion of both healthy and deviant individuals enabled estimation of both the FWER and power of methods. Standard deviations were computed by taking the square root of the respective diagonal element of the summed within-study and between-study covariance matrices. Both the norm data scores and the patient data scores were standardized, as required for the max-approach and stepwise approach.

A total of 1000 datasets (each consisting of one norm dataset and one patient dataset) were simulated for each type of data, enabling accurate estimation of FWER and power.

7.3.3 *Data Analysis*

For all methods, for each type of norm dataset, the FWER and power were estimated. The FWER was defined as the proportion of healthy patient datasets that were incorrectly identified as deviant – meaning that significant deviation on at least one test (at least one false positive result) was found (Huizenga et al., 2016). The significance threshold was set at $\alpha = 0.05$. The power was defined as the proportion of deviant patient datasets where deviation was correctly identified – meaning that deviation was found on the first test (Parikh, Mathai, Parikh, Sekhar, & Thomas, 2008). This definition of power was maintained regardless of the number of deviating tests.

7.4 RESULTS

Results were plotted for the default settings of 70 participants per study, from 20 studies, with 15 tests, with a between-tests correlation of 0.27, and a between-study variance of 0.15 (N = 70; S = 20; M = 15; BTC = 0.27; BSV = 0.15), unless otherwise noted. These settings were chosen to be typical for the ANDI database. Unless explicitly stated otherwise, no norm data were missing.

7.4.1 *Familywise Error Rate*

Our first question was whether multiple normative comparisons using a multilevel structured norm group required FWER correction. Figure 3 shows the FWER results for the typical ANDI settings. For uncorrected tests, the FWER results were well above 0.05, at approximately 0.40, confirming the necessity of using correction methods. All three correction methods kept the FWER at 0.05, suggesting adequate correction. Because the FWER of the uncorrected method was so high, this method will not be shown in later figures.

Second, FWER was plotted as a function of between-test correlation (BTC) and between-study variance (BSV), see Figure 4. This revealed that larger between-test correlations resulted in a minor decrease in the Bonferroni correction's FWER. Between-test correlations had no effect on FWER of the max-approach and stepwise approach. The between-study variance had a small effect on the FWER, where a high between-study variance increased the FWER to slightly above 0.05 across methods. The uncorrected method produced FWER values between 0.157 (BTC = 0.8; BSV = 0) and 0.567 (BTC = 0; BSV = 0.15).

Third, we looked at the influence of sample size on FWER. Sample size could either be changed by changing the number of studies (S), or by changing the number of participants per study (N). In Figure 5, different combinations of these two factors are shown. With a high sample size all three methods produced FWERs of 0.05, but increased FWER values were found as the sample size decreased; herein, decreasing the number of studies had a more pronounced effect than decreasing the number of participants per study. Noticeably,

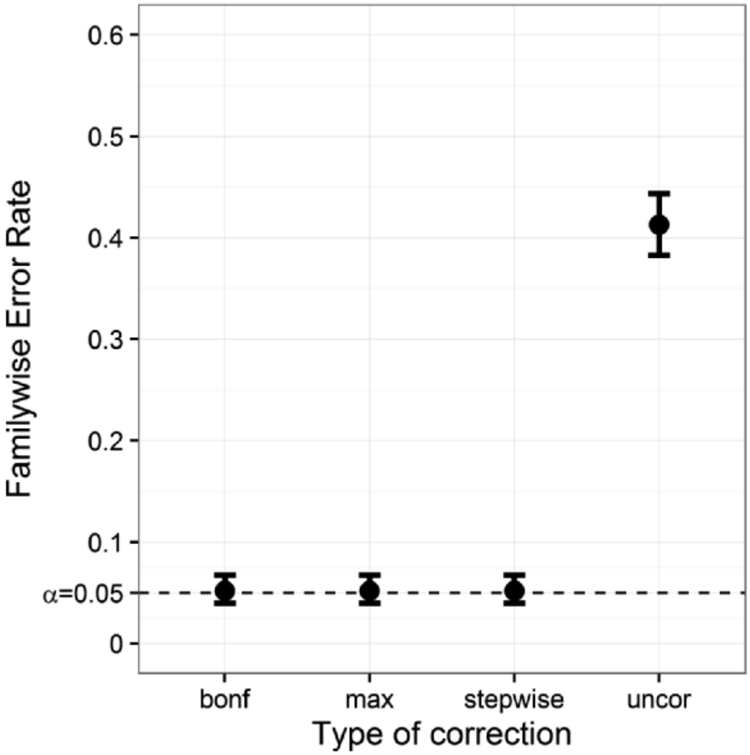

Figure 7.3: Familywise Error Rate on the y-axis, and type of correction on the x-axis. Plotted for the ANDI-representative settings (N = 70; S = 20; M = 15; BTC = 0.27; BSV = 0.15), without missing data. Error bars indicate 95% binomial confidence intervals. The dotted line indicates the significance threshold (α=0.05).

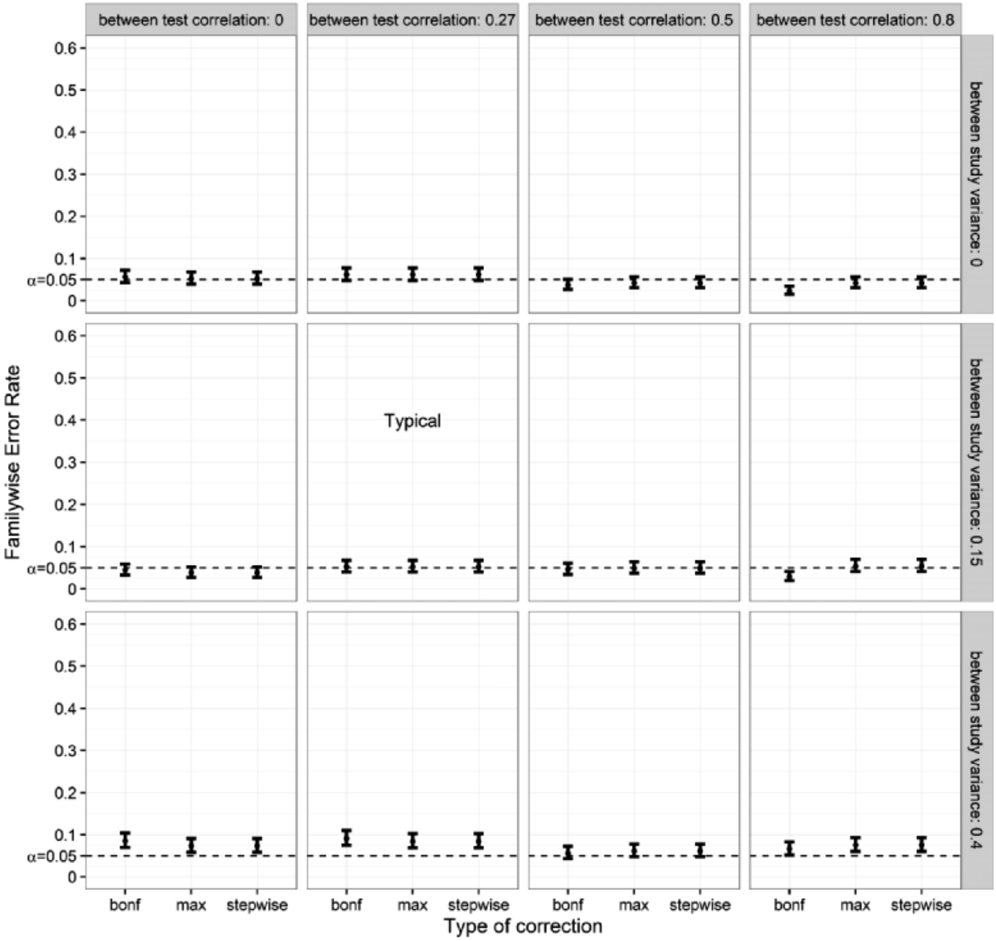

Figure 7.4: Familywise Error Rate on the y-axis, and type of correction on the x-axis. Plotted for various combinations of correlations between tests and various variances between studies (other parameters fixed at ANDI-representative settings: N = 70; S = 20; M = 15), without missing data. Error bars indicate 95% binomial confidence intervals. The dotted lines indicate the significance threshold (α=0.05). The graph marked by 'Typical' denotes that the between-test correlation and between-study variance corresponded to ANDI-representative settings (BTC = 0.27; BSV = 0.15).

the Bonferroni correction produced a higher FWER than the other two methods when the number of participants was low. The uncorrected method showed FWER values between 0.396 (S=40; N = 200) and 0.636 (S = 2; N = 10).

Fourth, we looked at the influence of the number of tests (M). The FWER of all correction methods for several numbers of test was plotted in Figure 6. For the Bonferroni correction, the FWER became elevated for 24 tests or more. The max-approach and stepwise approach showed no increased FWER. As expected, the uncorrected method showed a strong FWER increase as a result of multiple testing, with FWER = 0.055 (M = 1) to FWER = 0.719 (M = 50).

Fifth, we looked at the influence of missing data. Figure 7 displays the FWER of the three correction methods with either complete data or 50% of the data missing. Both the max-approach and the stepwise approach showed an increased FWER when missing data were introduced. The Bonferroni correction showed a negligibly small FWER increase. The uncorrected method appeared almost unaffected by missing data, with FWER = 0.42 (complete data) and FWER = 0.413 (missing data).

To summarize, FWER analysis revealed that the uncorrected method consistently produced FWER values above 0.05. This confirmed that performing multiple normative comparisons using multilevel data requires FWER correction. All correction methods produced better FWER values across a variety of situations. Between-test correlations slightly affected the FWER of the Bonferroni method, but not the FWER of the other correction methods. Between-study variance did affect FWER, with higher variances producing an increased FWER across correction methods, though only with relatively large between-study variances – which would be rare in clinical practice – and even then the increase was very mild. The number of tests only affected the Bonferroni correction, causing a small FWER increase as the number of tests increased. All correction methods showed an elevated FWER when the norm group was small, with the Bonferroni correction suffering most, especially when the number of participants was low. Missing data caused an increased FWER in the max-approach and stepwise approach alone. In short, the max-approach and stepwise approach outperformed the Bonferroni correction, especially when the norm data contained a low number of studies, or when the number of tests was high. Only when the norm data contained missing values, did the Bonferroni correction outperform the other correction methods.

7.4.2 *Power*

First, we looked at the power when the patient data deviated on the first test only, using the ANDI-representative settings (N = 70; S =

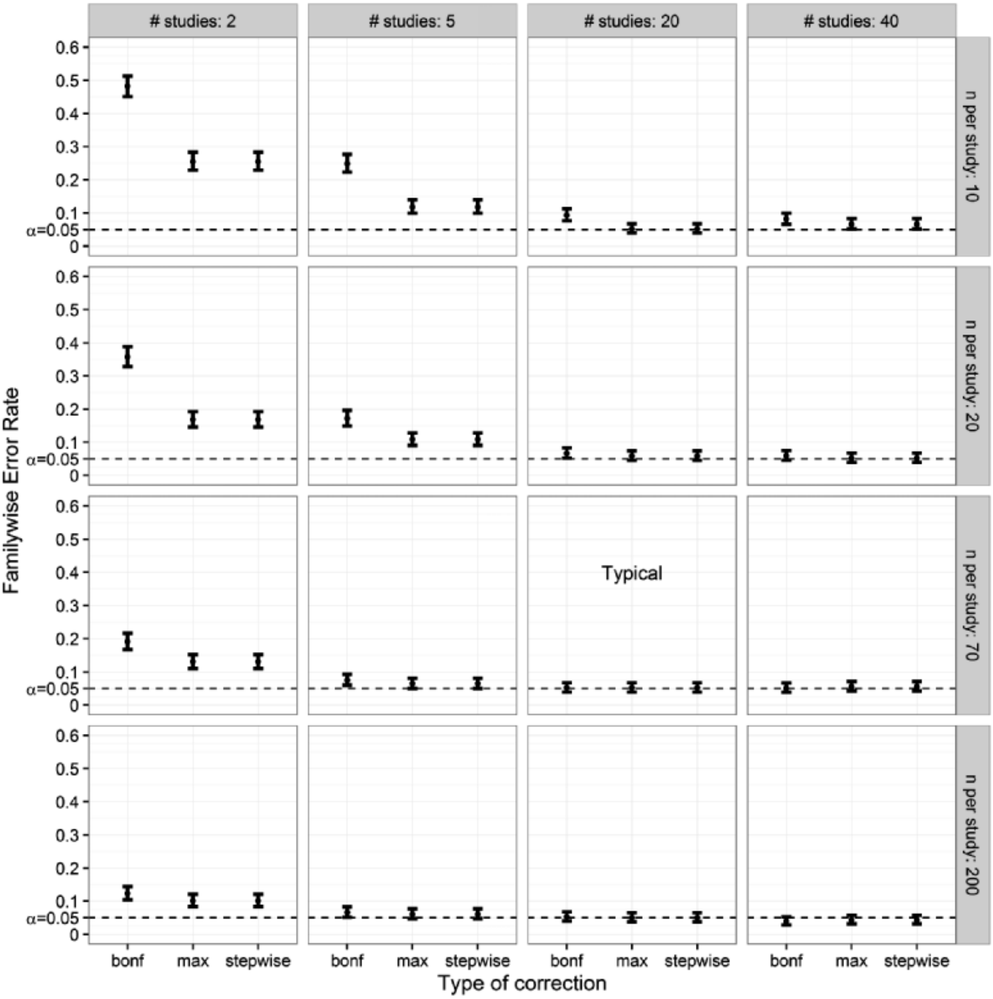

Figure 7.5: Familywise Error Rate on the y-axis, and type of correction on the x-axis. Plotted for various combinations of number of studies and number of participants per study (other parameters fixed at ANDI-representative settings: M = 15; BTC = 0.27; BSV = 0.15), without missing data. Error bars indicate 95% binomial confidence intervals. The dotted lines indicate the significance threshold (α=0.05). The graph marked by 'Typical' denotes that the number of studies and participants per study corresponded to ANDI-representative settings (S = 20; N = 70).

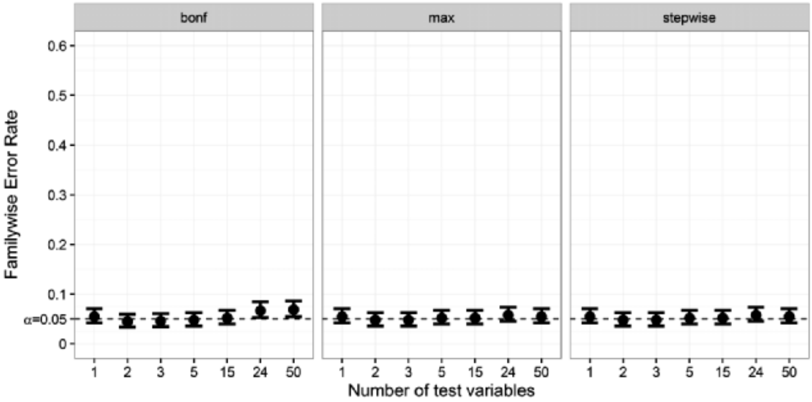

Figure 7.6: Familywise Error Rate on the y-axis, and number of tests on the x-axis. Plotted for various numbers of tests (other parameters fixed at ANDI-representative settings: N = 70; S = 20; BTC = 0.27; BSV = 0.15), without missing data. Error bars indicate 95% binomial confidence intervals. The dotted line indicates the significance threshold (α=0.05).

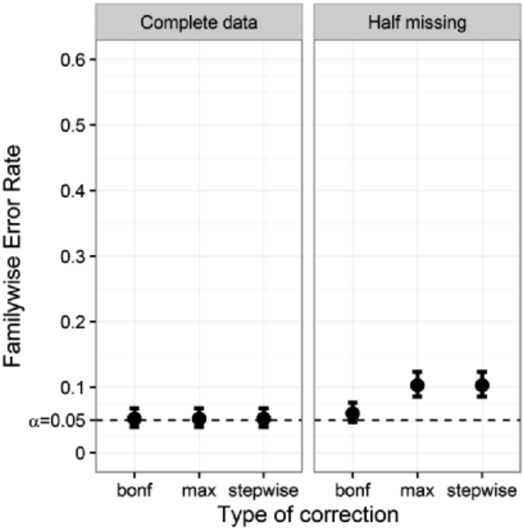

Figure 7.7: Familywise Error Rate on the y-axis, and type of correction on the x-axis. Plotted for both complete data (left) and data with half of the values removed (right), only for the ANDI-representative settings (N = 70; S = 20; M = 15; BTC = 0.27; BSV = 0.15). Error bars indicate 95% binomial confidence intervals. The dotted line indicates the significance threshold (α=0.05).

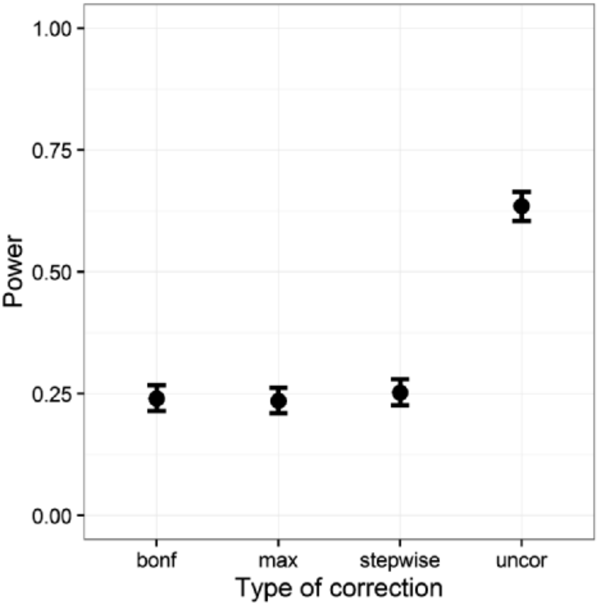

Figure 7.8: Power on the y-axis, and type of correction on the x-axis. Plotted for the ANDI-representative setting (N = 70; S = 20; M =15; BTC = 0.27; BSV = 0.15), without missing data. Error bars indicate 95% binomial confidence intervals.

20; M = 15; BTC = 0.27; BSV = 0.15). The power of the three correction methods and uncorrected normative comparison was plotted in Figure 8. The uncorrected method had the highest power. The three FWER correction methods produced almost identical results, and thus were concluded not to differ amongst each other.

Next, we looked at the power when the patient data deviated on the first five tests. Recall that power calculations only identified deviation on the first test. Figure 9 displays the power of all four methods while varying the correlations between tests. The uncorrected method still produced the highest power. Out of the correction methods, the stepwise approach had the highest power – even approximating the power of the uncorrected method, especially at low between-test correlations. The max-approach behaved in an opposite manner, showing increased power as between-test correlations increased, though never outperforming the stepwise approach. The Bonferroni method showed a consistently low power across between-test correlations.

To summarize, the uncorrected method produced the highest power. Unfortunately, this held little relevance as this method was already shown to fail in terms of FWER criteria. Out of the correction methods, the stepwise approach excelled when the assessed individual deviated on multiple tests. This agrees with the idea that both the

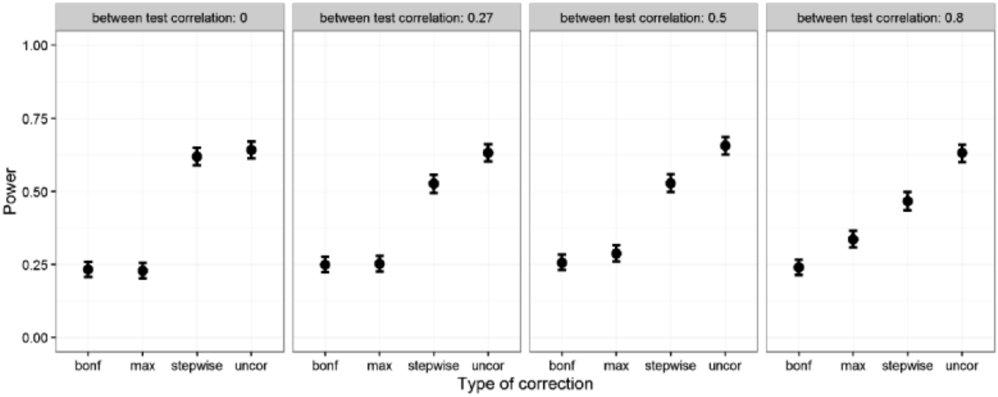

Figure 7.9: Power on the y-axis, and type of correction on the x-axis. Five deviations were simulated. Power was estimated as proportion of significant deviations found on the first test. Plotted for various between-tests correlations (other parameters fixed at ANDI-representative settings: N = 70; S = 20; M = 15; BSV = 0.15), without missing data. Error bars indicate 95% binomial confidence intervals.

Bonferroni and the max-approach are unfairly restrictive, especially for all except the most deviating test scores. Also, in neuropsychological assessment deviation on multiple tests is to be expected, as cognitive functions are correlated. As such, this advantage of the stepwise approach makes it very useful for clinical practice.

Other combinations of data simulation parameters that were varied are also available; all simulation results are provided online.

7.5 discussion

This study examined the assessment accuracy of several normative comparison methods when the norm group data were obtained from an aggregated dataset. The goal was to determine which method would be most suitable for use in neuropsychological assessment. Uncorrected normative comparison, and three FWER correction normative comparison methods – the Bonferroni correction, the max-approach, and the stepwise approach – were tested. Good assessment accuracy was defined as a familywise error rate (FWER) not exceeding the preset significance threshold. Additionally, the power was aimed to be as high as possible.

The uncorrected method consistently produced too high FWER values, meaning it too often untruthfully indicated that the assessed individual deviated from the norm group. The correction methods were shown to reduce the FWER. Several data parameters were varied to examine which correction method performed best under different circumstances. When the norm group contained many missing data, the Bonferroni correction controlled the FWER better than the max-approach and stepwise approach. Without missing data the stepwise approach performed preferably, as it had equivalent or better control over the FWER, and an equivalent or higher power across a variety of situations. This was especially pronounced in situations with a smaller number of studies or participants, situations with a higher number of tests, and when between-test correlations were low.

Several points require discussion. First, the max-approach and stepwise approach performed well as long as the norm group contained a sufficient amount of studies, while the Bonferroni correction suffered when either the number of studies or the number of participants was reduced. This difference can be explained by the fact that reducing norm group size results in fewer data points to make up the norm group distribution. Especially the tails of the distribution are affected by this, as they contain few data points to begin with. This affects the Bonferroni correction most because it implements a lower significance threshold for each test, and a lower threshold directs the comparison towards the most extreme part of the distribution (essentially the tail of the tail), which contains even fewer scores, and is thus even more affected by decreased sample size.

Second, introducing missing data to the norm group dataset led to an increased FWER in the max-approach and stepwise approach, but did not substantially affect the Bonferroni correction. This can be explained by the former two methods constructing norm group distributions by selecting extreme scores across tests; when half of the tests are missing these distributions may become too narrow (i.e. not critical enough). The Bonferroni correction isn't affected as it does not use the extreme values over all tests to make a new distribution to which the patient scores are compared.

Third, the stepwise approach produced a much higher power than the other correction methods when multiple tests deviated, especially when between-test correlations were low. This may be explained by the stepwise approach computing different distributions for each test score. More extreme scores are compared to more extreme distributions – distributions made up of the most extreme norm group scores. When tests are highly correlated, extreme scores on one test come with extreme scores on other tests, meaning there are more extreme scores in total. Thus the distributions become more critical, making it harder to detect deviation, thus reducing power.

Fourth, despite the stepwise approach yielding higher power than Bonferroni correction or max-approach, it occasionally produced a low power, which may spark reluctance to use it in clinical practice. However, the stepwise approach still outperformed the Bonferroni correction, and while the uncorrected method consistently produced the highest power, it also produced a highly increased FWER. It is the overall accuracy, the combination of a low FWER and relatively high power, that makes the stepwise approach most suitable for practical application. When high(er) power is preferred, we recommend a more liberal threshold (e.g. $\alpha = 0.20$ instead of $\alpha = 0.05$). This has the advantage of the true FWER being known (i.e. when the significance threshold of the stepwise approach is set to $\alpha = 0.20$, the resulting FWER will approximate 0.20), whereas using the uncorrected comparisons will produce an FWER increase of an unknown extent.

Fifth, norm data were simulated so that the number of participants was equal across studies, which is unlikely to occur in real aggregated data. A post-hoc simulation study with unequal sample sizes (using the default settings) showed similar patterns in terms of FWER results as it did with equal sample sizes.

Sixth, due to this being a simulation study, generalizability of results may be called into question. However, data simulation allowed for the examination of each method's performance under many different circumstances, thus boosting generalizability. More importantly, simulation parameters were based on real data to enhance generalizability, leading us to believe that these results are representative of real life situations.

Finally, it must be stressed that none of the discussed statistical methods are meant to be the sole basis of diagnosis, with contextual information and the assessors' professional opinion playing an important role – both in interpreting analysis results and in translating them into a meaningful judgement and effective treatment.

7.5.0.1 *Practical Advice*

When the norm data contain no missing data, the stepwise approach appears to be the most suitable method for normative comparison with an aggregated norm group; it best corrects the increased FWER associated with multiple testing, with FWER least affected by the properties of the norm group. Moreover, it produces a relatively high power when the assessed individual deviates on multiple tests. Based on this, we recommend the stepwise approach as the default method for neuropsychological assessment with an aggregated normative database. However, when the norm data contains (large portions of) missing data – for example, when several of the administered tests are relatively uncommon – the Bonferroni corrections should be preferred.

Also, when the norm group sample size is small, neither correction method performs well. In such instances, we recommend the resampling based normative comparison methods from Huizenga et al., (2016). These methods were made specifically with small norm groups in mind, and proved to have good assessment accuracy with small sample sizes (Huizenga et al., 2016; Li & Dye, 2013; Troendle, 1995). However, note that these methods have not yet been tested for multilevel data or norm groups with missing data.

The uncorrected method, the Bonferroni correction, the max-approach and stepwise approach have been implemented in a freely available online app (see: https://joost.shinyapps.io/EmpiricalNormComp/).

7.5.0.2 *Final Comments*

FWER corrections are needed in neuropsychological assessment when performing more than one normative comparison. In this simulation study, we have shown that correcting multiple comparisons using the stepwise approach can be a useful alternative to Bonferroni corrections when using aggregated norm data. We hope that this leads to a broader adoption of correction methods, as it is important to reduce the number of false positives in clinical practice, while remaining sensitive to true deviations.

8 STATISTICAL ADVANCES IN CLINICAL NEUROPSYCHOLOGY

8.1 SUMMARY

The goal of this thesis was to improve the reliability of neuropsychological assessment, specifically by improving the normative comparison procedure. The first goal was to provide multivariate normative comparisons, which test the patient's whole profile of scores. The second goal was to provide normative comparisons that are corrected for age, sex, and level of education. These goals had two requirements. First, a normative database had to be established with many, demographically diverse, healthy participants. Second, a statistical framework had to be developed that allows for demographically corrected multivariate normative comparisons with this new normative database. The statistical framework was the focus of this thesis.

In chapter two, we described how an aggregate normative database can be constructed by combining the data from healthy people from multiple studies. These people may have participated as a control group in a clinical study, or may have participated in a large community study. By combining many such groups of people, data from many different neuropsychological tests can be gathered. All procedures were standardized across tests. This involved two procedures for data cleaning. First, values were discarded that were outside a predefined range of allowable scores, which was set beforehand on the basis of clinical expertise. Second, values were discarded that were highly unlikely given participants' age, sex, and level of education. To select which demographic variables to use in demographic corrections, the Akaike Information Criterion was used. To be able to use parametric statistics, such as parametric normative comparisons, corrected scores would ideally be normally distributed, or transformed to be normally distributed. To select a power transformation to achieve normality, the Box-Cox procedure was used (Box & Cox, 1964). Last, the contents of the ANDI database were described in this chapter.

In chapter three, we described how multivariate normative comparisons can be made using an aggregate database. This required a model that consisted of three parts. First, to include demographic corrections for age, sex, and level of education, a regression model was required to estimate the regression coefficients for these three demographic variables. Second, there may be differences between studies in the scores that healthy participants obtain, for example due to differences in sample selection or test administration between studies. Therefore, a multilevel model was required, to model these differences between studies. Third, multivariate normative comparisons take into account the relations between scores on different tests. Therefore, the covariance between scores needed to be estimated, and a multivariate model was required. To combine these parts, a multivariate multilevel regression model was formulated. This multivariate multilevel regression has an added advantage, in that it can be fitted with missing data in the test variables. Because of the nature of an aggregate database, large amounts of missing data are to be expected, as tests that were not administered in a particular study have missing values for all participants in that study. With this model, all the components that are required in the multivariate normative comparisons can be estimated: demographically corrected means, variances, and covariances. In a simulation study, performance of the multivariate normative comparisons procedure was evaluated with varying amounts of missing data and between study variance. It was shown that although the model can be fitted using missing data, it cannot if there is missing overlap between tests. This issue was addressed in chapter four.

In chapter four, we described how the model from chapter three can be extended to accommodate missing overlap between tests. There is missing overlap between two tests, if the combination of these two tests has not been administered in any of the studies that are included in the database. This makes the covariance between these two tests impossible to estimate directly. In this chapter, two methods that can solve this problem are identified. The first is multiple imputation, where values are imputed for every missing value. From these imputed values, the covariance can be estimated in a straightforward manner. The second is a factor model approach, where a model for the covariance structure is estimated. This model assumes that the covariance between tests can be described by the dependence of these tests on the same latent variable. In a simulation study, the two methods are compared. The multiple imputation approach keeps the number of false positives under control, but due to underestimation of the covariance between tests, it is less sensitive in detecting impairment than the factor model approach. A precondition for the factor model approach is the appropriateness of the factor model for the data. If the factor model is not appropriate, the number of false positives increases. Therefore, a factor model for neuropsychological tests needs to be established before this model can be applied. This issue was addressed in chapter five.

In chapter five, the fit of different factor models for neuropsychological tests was compared in two studies. In the first study, a meta-analysis, correlation matrices for neuropsychological tests were requested from published studies. The correlation of test scores with demographic variables was partialed out from the correlation between tests. Subsequently, the correlation matrices were pooled into a single correlation matrix, to which factor models could be fitted. In the second study, factor models were fitted to demographically corrected data from the ANDI database. In both studies, model comparisons showed that the Cattell-Horn-Carroll model as modified by Jewsbury et al. (2016) fitted best. This model was originally developed in intelligence research, and divides cognitive functioning as measured by neuropsychological tests in domains of "Acquired knowledge or crystallized ability", "Processing speed", "Long-term memory encoding and retrieval", "Working memory", and "Word fluency". This is in contrast to other models that divide cognitive functioning in domains of "Attention", "Executive functioning", and "Memory". Because the Cattell-Horn-Carroll model seems to fit data from healthy people well, this model can be used in ANDI to apply the methods developed in chapter four.

In chapter six, the methods developed in this thesis were put to an empirical test. Specifically, the ANDI database and multivariate normative comparisons were used in a re-analysis of longitudinal data from a study on Parkinson's disease and Parkinson's disease dementia (Broeders et al., 2013). These data had been analyzed before using conventional (univariate) criteria for Mild Cognitive Impairment in Parkinson's disease (PD-MCI; Litvan et al., 2012). The goal of the previous study had been to see whether those who fit the PD-MCI criteria at the first measurement occasion would progress to dementia at a later measurement occasion. In this chapter, the results from this study were compared to results obtained with the ANDI database. First, using the univariate PD-MCI criteria with the ANDI database showed more cautious results than the earlier study: Fewer patients were classified as cognitively impaired. This was the case for both patients that later did, and did not develop dementia. Second, application of the ANDI database with multivariate normative comparisons was shown to provide better predictions than using the conventional PD-MCI criteria: They were both more sensitive and specific in predicting who would develop dementia. This provides evidence that the methods described here are indeed useful in improving neuropsychological assessment.

In chapter seven, we returned to the issue of using univariate normative comparisons in clinical neuropsychology. If no correction is used, and many univariate normative comparisons are performed for many different test variables, the number of times that cognition is judged to be impaired in healthy people is increased, a so-called increased familywise error rate. This may have contributed to the lower specificity for the PD-MCI criteria in chapter six. To correct for this increased familywise error rate, correction methods have been developed. A correction method that is frequently used in science, but not so much in clinical practice, is the Bonferroni correction. This correction decreases false positives, but can hurt the chances of detecting impairments in those who are truly impaired. In this study, more sophisticated correction methods were discussed and compared in a simulation study, specifically for the situation where patients are compared to an aggregate database. A new stepwise method performed better than the Bonferroni correction in detecting impairments in many settings, but did show an increase in false positives if many data were missing. Therefore, it is too early to fully endorse either method.

This thesis was accompanied by the establishment of the ANDI database and website. In this project, 84 generous contributions from research groups across the Netherlands and Belgium yielded data from 27,000 participants. In the ANDI project, the methods described in chapter two and three have been implemented. The website will be extended using the method from chapter four, with the model from chapter five.

8.2 POTENTIAL IMPROVEMENTS

The models in this dissertation were focused on multivariate distributions, and multilevel and factor model approaches. These approaches have had a number of advantages, in terms of the estimability of parameters in the light of many missing data points and differences between studies. However, these approaches brought with them a number of assumptions, in terms of linearity, equality of variances across different levels of demographic variables, and normality of the data. Voncken, Albers, & Timmerman (2017) proposed a powerful method for norming data that does not make these assumptions. However, their method is not multivariate, is not yet tested for very high percentages of missing values, and has not yet been developed for aggregate databases. One future direction could be to borrow the best from both methods, to arrive at multivariate normative comparisons while relaxing some of the assumptions where necessary.

The multivariate normative comparison procedure provides a single dichotomization into impaired and not impaired for a whole profile of test scores. This information is relevant in many clinical and research settings. However, in other settings, more detailed information on the nature of the deviation is needed, or a measure of the severity of impairment is needed. An option would be to study each of the test scores separately using univariate normative comparisons. Therefore, one approach would be to further improve the univariate approach from chapter seven, in order to make sure that it keeps the number of false positives low with missing normative data. Another approach would be to make multiple multivariate normative comparisons for parts of the profile, for example only comparing tests on two domains multivariately, ignoring the remainder of scores, and then comparing tests on two other domains multivariately. Whether this increases sensitivity to certain impairments, and how to control for false positives in this scenario are topics for future research.

The question remains how rare a patient's score or score profile has to be in healthy people before the patient's cognitive functioning can be classified as impaired. In this thesis, common thresholds were used, such as using the criterion that if a score this low is obtained by 5% of healthy people or less, we consider cognitive functioning to be impaired. How this threshold is set determines both sensitivity to real cognitive impairments, and the chance of finding a false positive. Therefore, it is important to use a threshold that maximizes performance in both respects. This 5% is therefore not set in stone, and should be considered a starting point. One reason that the best possible threshold has not been determined before, is that it is highly dependent on the context of the assessment. In some contexts, the base rate, i.e., the number of people with real cognitive impairments, will be higher than in other contexts. In a context with a high base rate, say at the intake of a memory clinic where patients with subjective complaints are invited, 5% may be too strict a threshold for impairment, and may result in many cognitive impairments being missed. In a context with a low base rate, say a screening of patients who have fallen recently, 5% may be too lenient a threshold for impairment, and may result in many false positives. Therefore, a fruitful extension of the present thesis would be to extend the multivariate normative comparisons method using Bayes' rule (Gavett, 2015) to take into account differences in base rates between different contexts.

8.3 extensions of the method

This thesis has focused on cross-sectional data from healthy people that completed a test battery for the first time. This makes this type of database useful to clinical neuropsychologists interested in setting diagnoses, and characterizing deficits in patients who complete a test battery for the first time. However, in for example treatment settings, clinical neuropsychologists also evaluate a patient's test scores at multiple occasions. To evaluate whether a patient's progression over time is different than observed in healthy people, an aggregate database of longitudinal data would ideally be built as well. There are multiple ways to envision such a longitudinal database.

One option would be to focus on a single retest session, for example after three months, for which normative data could be collected. Then, the decrease or increase of scores that patients show from baseline to this three-month follow-up can be compared to that of healthy people. This option could be difficult to implement if there are few studies that have retested healthy people after three months. Also, the application of this database would be limited to patients who are retested after three months. An advantage would be that the statistical framework of normative comparisons in this setting is already available in the form of the Reliable Change Index (Jacobson & Truax, 1991), which could also be extended to a multivariate version.

A second option would be to collect data from healthy people who were tested at varying time intervals. This would allow for the inclusion of more datasets, and could be more widely applicable because patients are tested at varying time intervals as well. However, the statistics for the modeling of changes over time would be more involved. From the longitudinal data, progressions of scores over time, i.e., the slope of the regression line, could be estimated for every healthy person. The patient's slope could then be compared to healthy participants' slopes in the by now familiar normative comparison procedures. Slopes of multiple test variables could then be analyzed using multivariate normative comparisons in the same way. One issue would be that time is not the only factor influencing differences between scores between measurement occasions, as practice effects from the measurement itself may improve scores over time (McCaffrey & Westervelt, 1995). This makes it difficult to pool data from a study with a measurement after three months, and a study with measurements every week.

Apart from longitudinal data, data from samples other than healthy people could be useful to clinical practice. This thesis was fundamentally focused on normative comparisons, that is, comparing a patient's scores to scores obtained by healthy people. An alternative would be to compare a patient's scores not only to scores obtained by healthy people, but also to scores obtained by clinical groups. With such data, a different statistical approach would be needed, to classify whether a patient's cognitive functioning is more similar to that of healthy people, or more similar to that of a particular clinical group. This would require that large samples of participants are available from different clinical groups, to best be able to make this distinction. We can be sure that studies are available amenable to the goal of pooling clinical groups, as in clinical studies often extensive test batteries are administered. Therefore, it would be worthwhile to add a database of patients from different studies to the existing database of healthy participants.

8.4 POTENTIAL APPLICATIONS

The focus of the project was on establishing a Dutch and Belgian normative database of tests used in clinical neuropsychology to diagnose cognitive impairment in adults. However, every step is generally applicable to normative comparisons in other domains. This is true for the pre-processing steps in chapter two, and is also true for the models of chapters three and four. It should be noted that the code required to perform these methods is available online, as is the code to run the website on which the normative comparisons can be made. Because of the generality of the methods proposed here, there are many applications that could follow naturally from this thesis.

First, the methods are not specific to Dutch and Belgian data. Therefore, this project can be replicated in other countries and/or regions. The availability of clinical neuropsychology norm data was already quite good in the Netherlands and Belgium, as test publishers provide high quality norms for this language area. Therefore, an aggregate database of normative data would be even more valuable in countries where there are fewer high quality norms available. One suggestion would be to keep the regions small for a single database: If studies from too many countries are aggregated, the variance between studies becomes large due to differences between countries in language, culture, and test versions.

Second, the methods are not specific to adults. Therefore, this project can be replicated for developmental clinical neuropsychology data as well. In fact, several donations to the ANDI database have already been made for children's data. It is not advisable to create a single database for children and adults, as the tests that are typically used are different between children and adults. Adding data from these age groups together would create a problem of missing overlap that is more severe than what was discussed in this thesis. Therefore, a new aggregate database of normative data for children would be advised.

Third, the methods developed here are not limited to applications in clinical neuropsychology. Clinical neuropsychology has the advantage that tests are highly standardized in the way they are administered, and in the way the outcomes of these tests are scored. This enables the combination of data across different sources. If there would be more flexibility in the administration procedure, this would contribute to more variance between studies. However, there are many fields within and outside of psychology where standardized tests and standardized outcomes are common, for example in clinical psychology (Clark & Watson, 1995), personnel psychology (Bartram, 2008), and educational science (Delandshere, 2001). For each of these fields, a normative database of test scores could be established, so multivariate normative comparisons can be used to classify whether a single case is typical, or different from the norm.

Fourth, although so far the discussion has only been about test scores, normative comparisons can be extended outside the domain of testing. Biomarkers would be one domain where an aggregate normative database would be useful. With such a normative database, for example a patient's blood pressure and heart rate variability could be compared to those of healthy people (Morrison & Morris, 1959; Tsuji et al., 1996). This could also be done in a multivariate normative comparison, taking the correlation between blood pressure and heart rate variability into account. Other biomarkers that would be amenable to normative comparisons could be brain indices (Dubois & Adolphs, 2016), obtained using fMRI, EEG or PET. Brain indices produce large amounts of data from different locations in the brain, which increases the chance of a false positive if every location is considered separately. Therefore, multivariate normative comparisons would be useful in this field as well.

8.5 THEMES

With this thesis and the broader ANDI project, we wanted to show that an aggregate normative database can provide clinical neuropsychologists with the data they need to apply statistically advanced techniques, which can improve diagnostics in clinical neuropsychology practice. Three themes that pervade this thesis are treated next.

The first theme is data sharing. The establishment of aggregate normative databases is only possible if there is broad willingness in the research community to share data. The ANDI project is very fortunate that there is a culture of cooperation in the clinical neuropsychology community in the Netherlands and Belgium. We hope that this will also be the case for future data aggregation ventures.

The second theme is integration of substantive and methodological fields within psychology. It generally takes a long time before newly developed statistical methods become available to other researchers, and before those methods that are available in research become available in clinical practice. With the advent of the free R statistics software with its productive community of developers, new statistical tools become freely available every day. However, the data and the know-how to apply these newly developed tools are also necessary. With the ANDI project, we hope to have crossed the divide between methodological development and substantive questions by developing a user-friendly website with which clinicians and researchers can analyze their own patient data.

The third and last theme is valorization. In recent years, there has been a call for science to become more useful to society at large. Scientists are asked to come up with studies that result in products that can be used, thereby providing value outside science. One criticism of this call for so-called valorization is that it could detract from research for which it is not immediately clear what the value is to society, but which may be valuable in its own right, or which may prove valuable in the long run. In the ANDI project, the data were not collected with valorization specifically in mind. However, they were re-used to create a product that is immediately useful to clinicians and patients. This seems like an ideal example of valorization.

8.6 conclusion

Neuropsychological assessment is an important part of clinical care and clinical research. When normative comparisons are reliable, we can use neuropsychological assessments to discover impairments in a patient's cognitive functioning, and to discover cognitive benefits and side effects of new treatments. Therefore, it is important to make sure that normative comparisons are as reliable as possible. In this thesis and the ANDI project, we improved normative comparisons, by developing a normative database and by developing a statistical framework to make normative comparisons with this database. These developments made multivariate normative comparisons, and more accurate demographic corrections, available to clinical neuropsychologists in practice and research.

9

APPENDICES ACCOMPANYING CHAPTER 5: COGNITIVE DOMAINS IN NEUROPSYCHOLOGY: SUPPORT FOR THE CATTELL-HORN-CARROLL MODEL IN TWO RESEARCH SYNTHESES

9.1 SEARCH TERMS USED IN PSYCINFO

#1 Factor model

factor analysis/ OR factor structure/ OR structural equation modeling/ OR (factor* model* OR factor* analy* OR structural equation* model* OR EFA OR CFA OR SEM OR factor* structur* OR confirmatory factor* OR exploratory factor*).ti,ab,id.

#2 Specific neuropsychological tests

stroop color word test/ OR stroop effect/ OR wechsler memory scale/ OR wisconsin card sorting test/ OR verbal learn*.tm. OR ((clock* AND (test* OR draw*)) OR (tower AND (test* OR london OR hanoi)) OR benton OR vis* retent* OR BVRT OR fac* recogni* OR BFRT OR judg* of line* OR line orientation OR JLO OR BJLO OR JOLO OR block design OR blockdesign OR Kohs OR boston naming OR BNT OR brixton OR spatial anticipation OR BSAT OR card sort* task* OR card sort* test* OR cardsort* test* OR WCST OR MWCST OR complex figur* OR rcf* OR rocf* OR rey-osterrieth OR digit* span* OR digitspan OR (span* ADJ1 (forward* OR back*)) OR spanforward OR digit* symbol* OR symbol* substitution* OR symbol coding OR DSST* OR family pictures OR figur* fluency OR groov* peg* OR purdue peg* OR pegboard OR letter fluency OR cowat OR controlled oral word association OR controlled word association OR controlled association* OR letter number OR LNS OR location learning OR LLT OR logical memory OR matr* reas* OR object* assemb* OR pac* audit* seri* additi* OR PASAT OR pict* arrangement* OR pict* compl* OR rivermead behavio* OR rbmt* OR selecti* remindi* OR srt OR Buschke OR VSRT OR semantic fluency OR verbal fluency OR category fluency OR animal* naming OR occupation* naming OR spatial span OR stroop OR symbol* search* OR trail making OR trial making OR tmt OR halstead reitan OR verbal learn* test* OR verbal learn* task* OR RAVLT* OR AVLT* OR CVLT* OR HVLT* OR verbal pair* associat* OR visual reproduction OR WMS*).ti,ab,id,tm.

#3 Clinical neuropsychological test batteries (test battery/ OR (((tests OR test scores OR test results) ADJ2 (attention* OR cognit* OR memory OR neuropsych* OR visual OR visuospatial* OR visuomotor OR verbal* OR executive OR learning OR IQ OR motor OR auditory OR perception OR inhibit* OR psychometr*)) OR (test* AND battery)).ti,ab,id,tm.) AND (neuropsychol*).ti,ab,id,hw,jx.

1 AND (2 OR 3)

9.2 TEST VARIABLES OF INTEREST.

Trail Making Test A, Trail Making Test B, Stroop Color, Stroop Word, Stroop Color-Word, Letter Fluency / FAS / COWAT, Semantic Fluency / Category Fluency / Animal Naming, Verbal Learning Test Total, Verbal Learning Test Recall, Verbal Learning Test Recognition, WAIS Vocabulary, WAIS Similarities, WAIS Information, WAIS Arithmetic, WAIS Letter Number Sequencing, WAIS Comprehension, WAIS Picture Completion, WAIS Block Design, WAIS Matrix Reasoning, WAIS Digit Symbol Substitution / Coding, WAIS Symbol Search, WAIS Picture Arrangement, WAIS Object Assembly, Logical Memory / Story Immediate, Logical Memory / Story Delayed, WMS Faces Immediate, WMS Faces Delayed, WMS Verbal Paired Associates Immediate, WMS Verbal Paired Associates Delayed, WMS Visual Paired Associates, WMS Family Pictures Immediate, WMS Family Pictures Delayed, WMS Visual Reproduction, WMS Spatial Span, Digit Span Forward, Digit Span Backward, Rey Complex Figure Copy, Rey Complex Figure Immediate Recall, Rey Complex Figure Delayed Recall, Raven Progressive Matrices, Wisconsin Number of Categories, Wisconsin Number of Perseverative Errors, Wisconsin Number of Perseverative Responses, Token Test Score, Grooved Pegboard Dominant, Grooved Pegboard Non-dominant, Benton Visual Retention Test, Brixton Spatial Anticipation, Rivermead Immediate 1 + 2, Rivermead Delayed 1 + 2, Clock Drawing Test, Boston Naming Test, Ruff Figural Fluency Test, Ruff 2 and 7, Buschke Selective Reminding Test Total Recall (TR), Buschke Selective Reminding Test Long Term Retrieval (LTR), Buschke Selective Reminding Test Long Term Storage (LTS), Buschke Selective Reminding Test Consistent Long Term Retrieval (CLTR), Free and Cued Selective Reminding Test (FCSRT), Buschke Selective Reminding Test Delayed Recall (DR), Benton Facial Recognition Test, Symbol Digit Modalities Test, Brief Visuospatial Memory Test, Judgement of Line Orientation, Tower of London Total number of moves, Continuous Performance Test (d′), Peabody Picture Vocabulary Test, PASAT Total number correct, BADS Zoo map, BADS Key search

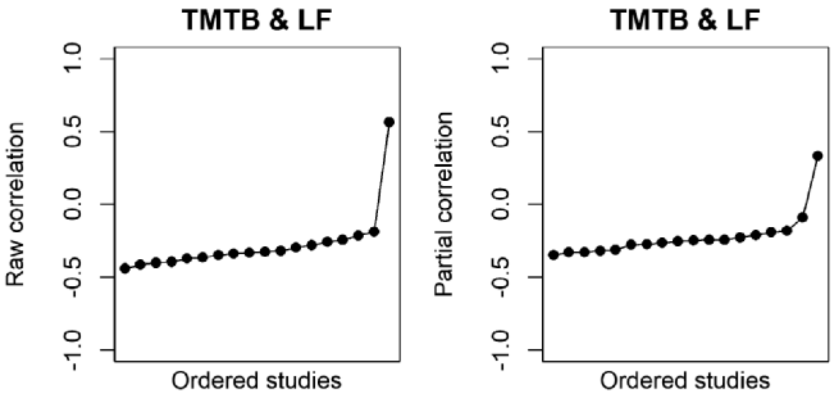

Figure 9.1: Bivariate raw and partial correlations between Trail Making Test B and Letter Fluency, plotted for different studies. The studies are ordered by the size of the correlation.

9.3 ANALYSIS WITHOUT TMT B FROM ROYALL ET AL. (2015)

Figure 1, left hand panel, shows that one correlation between TMTB and LF is exceptional, in that is positive and large. This is also the case for the correlation between TMTB and LMII from this study in Figure 2, so it is not LF that is at fault. These findings remain after partialing out the effect of age, sex, and level of education (right hand panel). This could be a case of a coding error, but Royall et al. (2015) is clear that the TMTB variable refers to the score in seconds, like other studies. Royall et al. (2015) also note that the correlations with TMTB seem strange. One last option is that it is simply due to sampling variance. However, given that this concerns an impressive 875 participants, this is unlikely. Other correlations that seemed different from the rest came from much smaller studies.

All correlations with TMTB were removed from the Royall correlation matrix for the main analysis, leaving LMII, BNT and LF. We did however run the analysis with these correlations with TMTB included. The results are given in Table 1. The conclusions do not differ from the conclusions of the main analysis: The second Jewsbury model was considered best in this analysis as well.

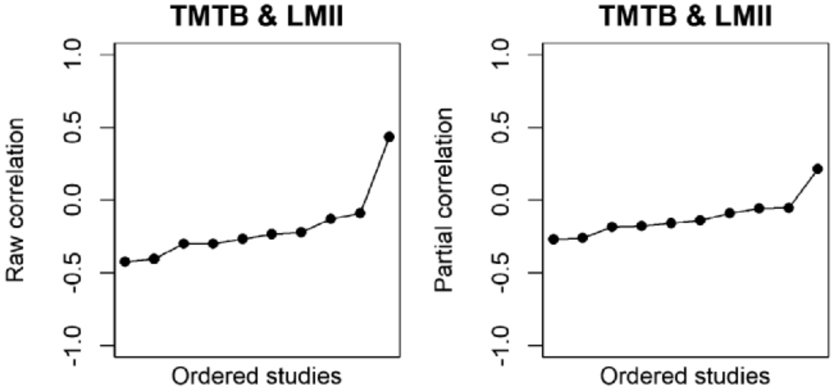

Figure 9.2: Bivariate raw and partial correlations between Trail Making Test B and Logical Memory II, plotted for different studies. The studies are ordered by the size of the correlation.

Table 9.1: Comparison Results with Correlations with TMT B from Royall et al. (2015) Included.

	χ^2(df)	RMSEA	SRMR	CFI	AIC	BIC
One factor	11193.5 (54)	0.058	0.218	0.937	11085.5	10599
Gross*	6698 (51)	0.046	0.147	0.962	6596	6136.5
Hoogland*	4672.1 (45)	0.041	0.118	0.974	4582.1	4176.7
Lezak	4886.3 (48)	0.041	0.122	0.973	4790.3	4357.9
Strauss	3828.7 (44)	0.038	0.112	0.979	3740.7	3344.4
Larrabee	3009.2 (48)	0.032	0.099	0.983	2913.2	2480.7
Jewsbury 1*	1347.6 (42)	0.023	0.058	0.993	1263.6	885.2
Jewsbury 2	1307.2 (41)	0.023	0.059	0.993	1225.2	855.9

*Model did not converge.

9.4 STUDY CHARACTERISTICS AND CORRELATION MATRICES

Table 9.2: Adrover-Roig, D., Sesé, A., Barceló, F., & Palmer, A. (2012). A latent variable approach to executive control in healthy ageing. *Brain and Cognition*, *78*(3), 284-299. doi:10.1016/j.bandc.2012.01.005

N = 122

Sex coding: male > female

Education coding: higher is better

Correlation matrix available from original publication

Table 9.3: Albert, M., Massaro, J., DeCarli, C., Beiser, A., Seshadri, S., Wolf, P. A., & Au, R. (2010). Profiles by sex of brain MRI and cognitive function in the framingham offspring study. *Alzheimer Disease and Associated Disorders*, *24*(2), 190-193. doi:10.1097/WAD.0b013e3181c1ed44

N = 2085

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LMI	LMII
AGE	1	-0.011	-0.218	0.311	0.398	-0.21	-0.227
SEX	-0.011	1	-0.096	-0.075	-0.038	0.11	0.12
EDU	-0.218	-0.096	1	-0.155	-0.286	0.311	0.307
TMTA	0.311	-0.075	-0.155	1	0.57	-0.206	-0.211
TMTB	0.398	-0.038	-0.286	0.57	1	-0.281	-0.299
LMI	-0.21	0.11	0.311	-0.206	-0.281	1	0.86
LMII	-0.227	0.12	0.307	-0.211	-0.299	0.86	1

Table 9.4: Andrejeva, N., Knebel, M., Dos Santos, V., Schmidt, J., Herold, C. J., Tudoran, R., ... & Gorenc-Mahmutaj, L. (2016). Neurocognitive deficits and effects of cognitive reserve in mild cognitive impairment. *Dementia and Geriatric Cognitive Disorders*, *41*(3-4), 199-209. doi:10.1159/000443791

N = 65

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LMI	LMII	SF	BNT	VLT-TR	VLT-DR
AGE	1	-0.121	-0.072	0.193	0.143	-0.241	-0.22	-0.156	-0.071	-0.128	-0.029
SEX	-0.121	1	-0.23	0.116	0.009	0.19	0.178	-0.036	-0.021	-0.357	-0.277
EDU	-0.072	-0.23	1	-0.198	-0.276	0.066	0.028	-0.104	0.201	0.08	0.042
TMTA	0.193	0.116	-0.198	1	0.403	0.012	-0.021	0.243	-0.011	-0.052	-0.183
TMTB	0.143	0.009	-0.276	0.403	1	-0.016	-0.091	0.259	0.092	0.2	0.063
LMI	-0.241	0.19	0.066	0.012	-0.016	1	0.864	0.083	0.274	0.069	0.043
LMII	-0.22	0.178	0.028	-0.021	-0.091	0.864	1	0.131	0.215	0.048	0.036
SF	-0.156	-0.036	-0.104	0.243	0.259	0.083	0.131	1	0.119	0.333	0.223
BNT	-0.071	-0.021	0.201	-0.011	0.092	0.274	0.215	0.119	1	0.145	0.061
VLT-TR	-0.128	-0.357	0.08	-0.052	0.2	0.069	0.048	0.333	0.145	1	0.567
VLT-DR	-0.029	-0.277	0.042	-0.183	0.063	0.043	0.036	0.223	0.061	0.567	1

Table 9.5: Andreotti, C., & Hawkins, K. A. (2015). RBANS norms based on the relationship of age, gender, education, and WRAT-3 reading to performance within an older African American sample. *The Clinical Neuropsychologist*, *29*(4), 442-465. doi:10.1080/13854046.2015.1039589

N = 289

Sex coding: Sex not included

Education coding: higher is better

Correlation matrix available from original publication

Table 9.6: Barnes, L. L., Yumoto, F., Capuano, A., Wilson, R. S., Bennett, D. A., & Tractenberg, R. E. (2016). Examination of the factor structure of a global cognitive function battery across race and time. *Journal of the International Neuropsychological Society*, *22*(1), 66-75. doi:10.1017/S1355617715001113

N = 2854

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	LMI	LMII	SF	DSF	DSB	COD	BNT	VLT-TR	VLT-DR
AGE	1	-0.013	-0.186	-0.258	-0.279	-0.321	-0.121	-0.097	-0.382	-0.185	-0.352	-0.339
SEX	-0.013	1	0.133	-0.064	-0.081	-0.115	0.042	-0.022	-0.077	0.065	-0.133	-0.126
EDU	-0.186	0.133	1	0.269	0.252	0.245	0.152	0.212	0.293	0.168	0.225	0.203
LMI	-0.258	-0.064	0.269	1	0.864	0.387	0.179	0.277	0.358	0.272	0.454	0.469
LMII	-0.279	-0.081	0.252	0.864	1	0.419	0.173	0.273	0.381	0.294	0.489	0.54
SF	-0.321	-0.115	0.245	0.387	0.419	1	0.203	0.29	0.498	0.337	0.491	0.475
DSF	-0.121	0.042	0.152	0.179	0.173	0.203	1	0.465	0.21	0.15	0.233	0.157
DSB	-0.097	-0.022	0.212	0.277	0.273	0.29	0.465	1	0.326	0.17	0.32	0.232
COD	-0.382	-0.077	0.293	0.358	0.381	0.498	0.21	0.326	1	0.381	0.424	0.407
BNT	-0.185	0.065	0.168	0.272	0.294	0.337	0.15	0.17	0.381	1	0.255	0.267
VLT-TR	-0.352	-0.133	0.225	0.454	0.489	0.491	0.233	0.32	0.424	0.255	1	0.727
VLT-DR	-0.339	-0.126	0.203	0.469	0.54	0.475	0.157	0.232	0.407	0.267	0.727	1

Table 9.7: Bennett, I. J., & Stark, C. E. (2016). Mnemonic discrimination relates to perforant path integrity: an ultra-high resolution diffusion tensor imaging study. *Neurobiology of Learning and Memory*, *129*, 107-112. doi:10.1016/j.nlm.2015.06.014

N = 109

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LMI	LMII	LF	SF	DSF	DSB	VLT-TR	VLT-DR
AGE	1	0.027	0.263	0.48	0.496	-0.292	-0.39	-0.002	-0.248	-0.235	-0.043	-0.32	-0.305
SEX	0.027	1	0.157	0.02	0.066	0.015	-0.029	0.106	0.19	0.153	0.221	-0.15	-0.155
EDU	0.263	0.157	1	0.1	0.002	0.044	0.086	0.085	0.082	-0.093	0.046	0.03	0.057
TMTA	0.48	0.02	0.1	1	0.718	-0.144	-0.211	-0.187	-0.375	-0.116	-0.136	-0.187	-0.143
TMTB	0.496	0.066	0.002	0.718	1	-0.34	-0.424	-0.215	-0.295	-0.268	-0.395	-0.338	-0.257
LMI	-0.292	0.015	0.044	-0.144	-0.34	1	0.877	0.084	0.325	0.253	0.34	0.467	0.523
LMII	-0.39	-0.029	0.086	-0.211	-0.424	0.877	1	0.166	0.293	0.235	0.339	0.51	0.621
LF	-0.002	0.106	0.085	-0.187	-0.215	0.084	0.166	1	0.223	0.252	0.319	0.173	0.097
SF	-0.248	0.19	0.082	-0.375	-0.295	0.325	0.293	0.223	1	0.156	0.22	0.2	0.146
DSF	-0.235	0.153	-0.093	-0.116	-0.268	0.253	0.235	0.252	0.156	1	0.384	0.319	0.196
DSB	-0.043	0.221	0.046	-0.136	-0.395	0.34	0.339	0.319	0.22	0.384	1	0.284	0.232
VLT-TR	-0.32	-0.15	0.03	-0.187	-0.338	0.467	0.51	0.173	0.2	0.319	0.284	1	0.769
VLT-DR	-0.305	-0.155	0.057	-0.143	-0.257	0.523	0.621	0.097	0.146	0.196	0.232	0.769	1

Table 9.8: Bezdicek, O., Libon, D. J., Stepankova, H., Panenkova, E., Lukavsky, J., Garrett, K. D., ... & Kopecek, M. (2014). Development, validity, and normative data study for the 12-word Philadelphia Verbal Learning Test [czP (r) VLT-12] among older and very old Czech adults. *The Clinical Neuropsychologist*, *28*(7), 1162-1181. doi:10.1080/13854046.2014.952666

N = 540

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTB	LMI	LMII	LF	SF	DSF	DSB	BNT
AGE	1	0.012	-0.154	0.39	-0.177	-0.228	-0.171	-0.359	-0.213	-0.225	-0.264
SEX	0.012	1	-0.164	0.007	-0.121	-0.108	0.044	0.063	0.022	-0.087	-0.234
EDU	-0.154	-0.164	1	-0.296	0.244	0.273	0.246	0.357	0.304	0.277	0.275
TMTB	0.39	0.007	-0.296	1	-0.157	-0.22	-0.331	-0.444	-0.257	-0.271	-0.335
LMI	-0.177	-0.121	0.244	-0.157	1	0.87	0.262	0.224	0.153	0.301	0.419
LMII	-0.228	-0.108	0.273	-0.22	0.87	1	0.287	0.308	0.189	0.284	0.432
LF	-0.171	0.044	0.246	-0.331	0.262	0.287	1	0.541	0.3	0.349	0.331
SF	-0.359	0.063	0.357	-0.444	0.224	0.308	0.541	1	0.305	0.322	0.368
DSF	-0.213	0.022	0.304	-0.257	0.153	0.189	0.3	0.305	1	0.493	0.168
DSB	-0.225	-0.087	0.277	-0.271	0.301	0.284	0.349	0.322	0.493	1	0.3
BNT	-0.264	-0.234	0.275	-0.335	0.419	0.432	0.331	0.368	0.168	0.3	1

Table 9.9: Booth, T., Royle, N. A., Corley, J., Gow, A. J., Hernández, M. D. C. V., Maniega, S. M., ... & Deary, I. J. (2015). Association of allostatic load with brain structure and cognitive ability in later life. *Neurobiology of Aging*, *36*(3), 1390-1399. doi:10.1016/j.neurobiolaging.2014.12.020

N = 970

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	LMI	LMII	SF	DSB	COD
AGE	1	0.02	-0.071	-0.18	-0.169	-0.144	-0.141	-0.198
SEX	0.02	1	-0.029	0.077	0.108	0.059	-0.041	0.162
EDU	-0.071	-0.029	1	0.307	0.287	0.241	0.199	0.285
LMI	-0.18	0.077	0.307	1	0.873	0.197	0.238	0.238
LMII	-0.169	0.108	0.287	0.873	1	0.195	0.23	0.244
SF	-0.144	0.059	0.241	0.197	0.195	1	0.28	0.342
DSB	-0.141	-0.041	0.199	0.238	0.23	0.28	1	0.264
COD	-0.198	0.162	0.285	0.238	0.244	0.342	0.264	1

Table 9.10: Bouazzaoui, B., Fay, S., Taconnat, L., Angel, L., Vanneste, S., & Isingrini, M. (2013). Differential involvement of knowledge representation and executive control in episodic memory performance in young and older adults. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *67*(2), 100-107. doi:10.1037/a0028517

N = 120

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	LF	SF
AGE	1	0.093	-0.229	-0.324	-0.23
SEX	0.093	1	-0.099	-0.189	-0.067
EDU	-0.229	-0.099	1	0.312	0.038
LF	-0.324	-0.189	0.312	1	0.411
SF	-0.23	-0.067	0.038	0.411	1

Table 9.11: Bowden, S. C., Cook, M. J., Bardenhagen, F. J., Shores, E. A., & Carstairs, J. R. (2004). Measurement invariance of core cognitive abilities in heterogeneous neurological and community samples. *Intelligence*, *32*(4), 363-389. doi:10.1016/j.intell.2004.05.002

N = 399

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	LMI	LMII	DSF	DSB	COD	VLT-TR	VLT-DR
AGE	1	0.017	0.018	0.002	0.009	-0.082	-0.025	-0.148	-0.025	-0.045
SEX	0.017	1	-0.028	0.16	0.164	-0.076	-0.01	0.315	0.248	0.226
EDU	0.018	-0.028	1	0.189	0.188	0.132	0.182	0.238	0.278	0.176
LMI	0.002	0.16	0.189	1	0.916	0.133	0.21	0.252	0.533	0.468
LMII	0.009	0.164	0.188	0.916	1	0.121	0.196	0.238	0.531	0.51
DSF	-0.082	-0.076	0.132	0.133	0.121	1	0.57	0.167	0.128	0.035
DSB	-0.025	-0.01	0.182	0.21	0.196	0.57	1	0.247	0.289	0.173
COD	-0.148	0.315	0.238	0.252	0.238	0.167	0.247	1	0.332	0.296
VLT-TR	-0.025	0.248	0.278	0.533	0.531	0.128	0.289	0.332	1	0.745
VLT-DR	-0.045	0.226	0.176	0.468	0.51	0.035	0.173	0.296	0.745	1

Table 9.12: Bunce, D., Batterham, P. J., Christensen, H., & Mackinnon, A. J. (2014). Causal associations between depression symptoms and cognition in a community-based cohort of older adults. *The American Journal of Geriatric Psychiatry*, *22*(12), 1583-1591. doi:10.1016/j.jagp.2014.01.004

N = 853

Sex coding: male > female

Education coding: higher is better

AGE	1	-0.07	-0.07	-0.24	-0.33
SEX	-0.07	1	0.17	0.07	0.02
EDU	-0.07	0.17	1	0.16	0.34
SF	-0.24	0.07	0.16	1	0.47
COD	-0.33	0.02	0.34	0.47	1

Table 9.13: Chan, R. C., Wang, Y., Wang, L., Chen, E. Y., Manschreck, T. C., Li, Z. J., ... & Gong, Q. Y. (2009). Neurological soft signs and their relationships to neurocognitive functions: A re-visit with the structural equation modeling design. *PLoS One*, *4*(12), 1-8. doi:10.1371/journal.pone.0008469

N = 160

Sex coding: male > female

Education coding: higher is better

Table 9.14: Chen, Y. C., Jung, C. C., Chen, J. H., Chiou, J. M., Chen, T. F., Chen, Y. F., ... & Lee, M. S. (2017). Association of dietary patterns with global and domain-specific cognitive decline in Chinese elderly. *Journal of the American Geriatrics Society*, *65*(6), 1159-1167. doi:10.1111/jgs.14741

N = 475

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LMI	LMII	SF	DSB
AGE	1	0.238	-0.089	0.376	0.424	-0.336	-0.301	-0.349	-0.257
SEX	0.238	1	0.287	-0.038	0.007	-0.027	-0.023	-0.36	-0.036
EDU	-0.089	0.287	1	-0.373	-0.25	0.302	0.324	-0.017	0.289
TMTA	0.376	-0.038	-0.373	1	0.529	-0.297	-0.302	-0.292	-0.311
TMTB	0.424	0.007	-0.25	0.529	1	-0.348	-0.3	-0.257	-0.259
LMI	-0.336	-0.027	0.302	-0.297	-0.348	1	0.89	0.358	0.372
LMII	-0.301	-0.023	0.324	-0.302	-0.3	0.89	1	0.332	0.362
SF	-0.349	-0.36	-0.017	-0.292	-0.257	0.358	0.332	1	0.238
DSB	-0.257	-0.036	0.289	-0.311	-0.259	0.372	0.362	0.238	1

Table 9.15: Ciccarelli, N., Fabbiani, M., Baldonero, E., Fanti, I., Cauda, R., Giambenedetto, S. D., & Silveri, M. C. (2012). Effect of aging and human immunodeficiency virus infection on cognitive abilities. *Journal of the American Geriatrics Society*, *60*(11), 2048-2055. doi:10.1111/j.1532-5415.2012.04213.x

N = 39

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	TMTB	LF	DSF	DSB	COD	VLT-TR	VLT-DR
AGE	1	-0.112	-0.305	0.664	-0.314	-0.4	-0.657	-0.458	-0.532	-0.441
SEX	-0.112	1	0.238	0.064	-0.02	0.169	0.172	0.38	-0.201	-0.218
EDU	-0.305	0.238	1	-0.339	0.368	0.531	0.466	0.191	0.138	0.259
TMTB	0.664	0.064	-0.339	1	-0.415	-0.299	-0.555	-0.641	-0.404	-0.47
LF	-0.314	-0.02	0.368	-0.415	1	0.326	0.442	0.102	0.444	0.536
DSF	-0.4	0.169	0.531	-0.299	0.326	1	0.563	0.299	0.238	0.185
DSB	-0.657	0.172	0.466	-0.555	0.442	0.563	1	0.349	0.454	0.384
COD	-0.458	0.38	0.191	-0.641	0.102	0.299	0.349	1	0.184	0.244
VLT-TR	-0.532	-0.201	0.138	-0.404	0.444	0.238	0.454	0.184	1	0.785
VLT-DR	-0.441	-0.218	0.259	-0.47	0.536	0.185	0.384	0.244	0.785	1

Table 9.16: Darst, B. F., Koscik, R. L., Hermann, B. P., La Rue, A., Sager, M. A., Johnson, S. C., & Engelman, C. D. (2015). Heritability of cognitive traits among siblings with a parental history of Alzheimer's disease. *Journal of Alzheimer's Disease*, *45*(4), 1149-1155. doi:10.3233/JAD-142658

N = 1226

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LMI	LMII	DSF	DSB	VLT-TR	VLT-DR
AGE	1	-0.042	0.055	0.312	0.312	-0.09	-0.11	-0.048	-0.023	-0.225	-0.161
SEX	-0.042	1	-0.073	-0.043	-0.024	-0.029	-0.026	-0.046	-0.017	0.157	0.147
EDU	0.055	-0.073	1	0.019	-0.12	0.238	0.237	0.166	0.215	0.136	0.148
TMTA	0.312	-0.043	0.019	1	0.496	-0.091	-0.11	-0.107	-0.13	-0.239	-0.18
TMTB	0.312	-0.024	-0.12	0.496	1	-0.222	-0.234	-0.219	-0.255	-0.275	-0.214
LMI	-0.09	-0.029	0.238	-0.091	-0.222	1	0.897	0.166	0.266	0.442	0.407
LMII	-0.11	-0.026	0.237	-0.11	-0.234	0.897	1	0.13	0.232	0.461	0.467
DSF	-0.048	-0.046	0.166	-0.107	-0.219	0.166	0.13	1	0.562	0.17	0.056
DSB	-0.023	-0.017	0.215	-0.13	-0.255	0.266	0.232	0.562	1	0.24	0.16
VLT-TR	-0.225	0.157	0.136	-0.239	-0.275	0.442	0.461	0.17	0.24	1	0.766
VLT-DR	-0.161	0.147	0.148	-0.18	-0.214	0.407	0.467	0.056	0.16	0.766	1

Table 9.17: Duff, K. D., Langbehn, D. R., Schoenberg, M. R., Moser, D. J., Baade, L. E., Mold, J. W., . . . Adams, R. L. (2006). Examining the repeatable battery for the assessment of neuropsychological status: Factor analytic studies in an elderly sample. *The American Journal of Geriatric Psychiatry, 14*, 976-979. doi:10.1097/01.JGP.0000229690.70011

N = 823

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	LMI	LMII	SF	DSF	COD	BNT	VLT-TR	VLT-DR
AGE	1	-0.062	-0.007	-0.205	-0.247	-0.191	-0.059	-0.424	-0.155	-0.249	-0.23
SEX	-0.062	1	0.18	0.016	-0.053	-0.178	0.09	-0.07	0.183	-0.1	-0.171
EDU	-0.007	0.18	1	0.248	0.209	0.085	0.187	0.221	0.275	0.192	0.108
LMI	-0.205	0.016	0.248	1	0.787	0.343	0.291	0.383	0.313	0.582	0.55
LMII	-0.247	-0.053	0.209	0.787	1	0.344	0.206	0.424	0.318	0.575	0.623
SF	-0.191	-0.178	0.085	0.343	0.344	1	0.168	0.416	0.225	0.393	0.371
DSF	-0.059	0.09	0.187	0.291	0.206	0.168	1	0.213	0.133	0.275	0.113
COD	-0.424	-0.07	0.221	0.383	0.424	0.416	0.213	1	0.356	0.433	0.377
BNT	-0.155	0.183	0.275	0.313	0.318	0.225	0.133	0.356	1	0.243	0.232
VLT-TR	-0.249	-0.1	0.192	0.582	0.575	0.393	0.275	0.433	0.243	1	0.65
VLT-DR	-0.23	-0.171	0.108	0.55	0.623	0.371	0.113	0.377	0.232	0.65	1

Table 9.18: Eifler, S., Rausch, F., Schirmbeck, F., Veckenstedt, R., Englisch, S., Meyer-Lindenberg, A., ... & Zink, M. (2014). Neurocognitive capabilities modulate the integration of evidence in schizophrenia. *Psychiatry Research*, *219*(1), 72-78. doi:10.1016/j.psychres.2014.04.056

N = 52

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	SF	COD	VLT-TR
AGE	1	-0.104	0.436	-0.038	-0.053	0.427	-0.118	0.099
SEX	-0.104	1	-0.025	-0.052	-0.057	0.042	0.368	0.085
EDU	0.436	-0.025	1	-0.225	-0.168	0.301	0.153	0.318
TMTA	-0.038	-0.052	-0.225	1	0.486	-0.314	-0.353	-0.043
TMTB	-0.053	-0.057	-0.168	0.486	1	-0.351	-0.269	-0.143
SF	0.427	0.042	0.301	-0.314	-0.351	1	0.093	0.171
COD	-0.118	0.368	0.153	-0.353	-0.269	0.093	1	0.325
VLT-TR	0.099	0.085	0.318	-0.043	-0.143	0.171	0.325	1

Table 9.19: Fernaeus, S. E., Östberg, P., Wahlund, L. O., & Hellström, Å. (2014). Memory factors in Rey AVLT: implications for early staging of cognitive decline. *Scandinavian Journal of Psychology*, *55*(6), 546-553. doi:10.1111/sjop.12157

N = 42

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	TMTA	LMI	LMII	DSF	DSB	VLT-TR
AGE	1	-0.011	-0.407	-0.028	-0.239	-0.443	-0.241	-0.161	-0.367
SEX	-0.011	1	-0.286	0.222	-0.111	-0.282	0.154	0.151	-0.217
EDU	-0.407	-0.286	1	-0.264	0.3	0.187	0.058	0.088	0.476
TMTA	-0.028	0.222	-0.264	1	-0.367	-0.3	-0.289	-0.253	-0.202
LMI	-0.239	-0.111	0.3	-0.367	1	0.339	0.06	0.067	0.387
LMII	-0.443	-0.282	0.187	-0.3	0.339	1	-0.03	-0.062	0.071
DSF	-0.241	0.154	0.058	-0.289	0.06	-0.03	1	0.695	0.348
DSB	-0.161	0.151	0.088	-0.253	0.067	-0.062	0.695	1	0.435
VLT-TR	-0.367	-0.217	0.476	-0.202	0.387	0.071	0.348	0.435	1

Table 9.20: Ferreira, N. V., Cunha, P. J., da Costa, D. I., dos Santos, F., Costa, F. O., Consolim-Colombo, F., & Irigoyen, M. C. (2015). Association between functional performance and executive cognitive functions in an elderly population including patients with low ankle–brachial index. *Clinical Interventions in Aging*, *10*, 839-847. doi:10.2147/CIA.S69270

N = 40

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	LF	SF	DSF	DSB
AGE	1	-0.264	0.213	-0.054	-0.066	0.058	-0.266
SEX	-0.264	1	-0.139	0.152	0.287	0.262	0.356
EDU	0.213	-0.139	1	0.243	0.28	0.177	-0.024
LF	-0.054	0.152	0.243	1	0.539	0.24	0.48
SF	-0.066	0.287	0.28	0.539	1	0.42	0.501
DSF	0.058	0.262	0.177	0.24	0.42	1	0.508
DSB	-0.266	0.356	-0.024	0.48	0.501	0.508	1

Table 9.21: Fortin, A., & Caza, N. (2014). A validation study of memory and executive functions indexes in French-speaking healthy young and older adults. *Canadian Journal on Aging/La Revue canadienne du vieillissement*, *33*(1), 60-71. doi:10.1017/S0714980813000445

N = 98

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	LMI	LF	DSB	VLT-DR
AGE	1	0.027	0.005	-0.408	0.11	-0.382	-0.315
SEX	0.027	1	-0.234	0.062	-0.026	-0.006	0.344
EDU	0.005	-0.234	1	0.045	0.279	0.074	0.072
LMI	-0.408	0.062	0.045	1	-0.114	0.16	0.4
LF	0.11	-0.026	0.279	-0.114	1	0.13	0.117
DSB	-0.382	-0.006	0.074	0.16	0.13	1	0.253
VLT-DR	-0.315	0.344	0.072	0.4	0.117	0.253	1

Table 9.22: Gallagher, P., Gray, J. M., Watson, S., Young, A. H., & Ferrier, I. N. (2014). Neurocognitive functioning in bipolar depression: a component structure analysis. *Psychological Medicine*, *44*(5), 961-974. doi:10.1017/S0033291713001487

N = 47

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	LF	DSF	DSB	COD	VLT-TR	VLT-DR
AGE	1	0.261	-0.097	0.196	-0.102	-0.025	-0.191	-0.442	-0.359
SEX	0.261	1	-0.046	0.196	0.037	-0.056	0.383	-0.007	-0.1
EDU	-0.097	-0.046	1	0.127	0.029	0.383	0.397	0.247	0.15
LF	0.196	0.196	0.127	1	0.11	0.441	0.18	0.083	0.092
DSF	-0.102	0.037	0.029	0.11	1	0.244	0.275	0.232	0.189
DSB	-0.025	-0.056	0.383	0.441	0.244	1	0.249	0.05	0.192
COD	-0.191	0.383	0.397	0.18	0.275	0.249	1	0.312	0.316
VLT-TR	-0.442	-0.007	0.247	0.083	0.232	0.05	0.312	1	0.817
VLT-DR	-0.359	-0.1	0.15	0.092	0.189	0.192	0.316	0.817	1

Table 9.23: Horvat, P., Richards, M., Malyutina, S., Pajak, A., Kubinova, R., Tamosiunas, A., ... & Bobak, M. (2014). Life course socioeconomic position and mid-late life cognitive function in Eastern Europe. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *69*(3), 470-481. doi:10.1093/geronb/gbu014

Country: Czech Republic

N = 5490

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	SF	VLT-TR
AGE	1	-0.08	-0.09	-0.21	-0.26
SEX	-0.08	1	-0.32	0.01	0.27
EDU	-0.09	-0.32	1	0.3	0.3
SF	-0.21	0.01	0.3	1	0.4
VLT-TR	-0.26	0.27	0.3	0.4	1

Table 9.24: Horvat, P., Richards, M., Malyutina, S., Pajak, A., Kubinova, R., Tamosiunas, A., ... & Bobak, M. (2014). Life course socioeconomic position and mid-late life cognitive function in Eastern Europe. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *69*(3), 470-481. doi:10.1093/geronb/gbu014

Country: Lithuania

N = 6762

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	SF	VLT-TR
AGE	1	-0.02	-0.21	-0.29	-0.37
SEX	-0.02	1	-0.01	-0.01	0.28
EDU	-0.21	-0.01	1	0.4	0.43
SF	-0.29	-0.01	0.4	1	0.4
VLT-TR	-0.37	0.28	0.43	0.4	1

Table 9.25: Horvat, P., Richards, M., Malyutina, S., Pajak, A., Kubinova, R., Tamosiunas, A., ... & Bobak, M. (2014). Life course socioeconomic position and mid-late life cognitive function in Eastern Europe. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *69*(3), 470-481. doi:10.1093/geronb/gbu014

Country: Poland

N = 10317

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	SF	VLT-TR
AGE	1	-0.05	-0.11	-0.29	-0.36
SEX	-0.05	1	-0.09	0	0.16
EDU	-0.11	-0.09	1	0.38	0.36
SF	-0.29	0	0.38	1	0.55
VLT-TR	-0.36	0.16	0.36	0.55	1

Table 9.26: Horvat, P., Richards, M., Malyutina, S., Pajak, A., Kubinova, R., Tamosiunas, A., ... & Bobak, M. (2014). Life course socioeconomic position and mid-late life cognitive function in Eastern Europe. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *69*(3), 470-481. doi:10.1093/geronb/gbu014

Country: Russia

N = 8277

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	SF	VLT-TR
AGE	1	-0.02	-0.17	-0.38	-0.42
SEX	-0.02	1	-0.04	-0.03	0.17
EDU	-0.17	-0.04	1	0.28	0.29
SF	-0.38	-0.03	0.28	1	0.47
VLT-TR	-0.42	0.17	0.29	0.47	1

Table 9.27: Hedden, T., & Yoon, C. (2006). Individual differences in executive processing predict susceptibility to interference in verbal working memory. *Neuropsychology*, *20*(5), 511-528. doi:10.1037/0894-4105.20.5.511.supp.

N = 121

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	DSB
AGE	1	-0.16	0.14	0.34	0.34	-0.19
SEX	-0.16	1	-0.21	-0.07	0.01	-0.07
EDU	0.14	-0.21	1	0.05	-0.11	0.25
TMTA	0.34	-0.07	0.05	1	0.57	-0.23
TMTB	0.34	0.01	-0.11	0.57	1	-0.39
DSB	-0.19	-0.07	0.25	-0.23	-0.39	1

Table 9.28: Hedden, T., Mormino, E. C., Amariglio, R. E., Younger, A. P., Schultz, A. P., Becker, J. A., ... & Rentz, D. M. (2012). Cognitive profile of amyloid burden and white matter hyperintensities in cognitively normal older adults. *Journal of Neuroscience*, *32*(46), 16233-16242. doi:10.1523/JNEUROSCI.2462-12.2012

N = 168

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LF	SF	DSB	COD
AGE	1	-0.04	-0.05	0.12	0.22	0.01	-0.19	-0.03	-0.22
SEX	-0.04	1	-0.09	0.07	0.14	0.03	0.06	-0.15	0.06
EDU	-0.05	-0.09	1	-0.22	-0.38	0.35	0.36	0.3	0.3
TMTA	0.12	0.07	-0.22	1	0.46	-0.13	-0.3	-0.12	-0.53
TMTB	0.22	0.14	-0.38	0.46	1	-0.35	-0.4	-0.28	-0.53
LF	0.01	0.03	0.35	-0.13	-0.35	1	0.56	0.35	0.38
SF	-0.19	0.06	0.36	-0.3	-0.4	0.56	1	0.33	0.48
DSB	-0.03	-0.15	0.3	-0.12	-0.28	0.35	0.33	1	0.28
COD	-0.22	0.06	0.3	-0.53	-0.53	0.38	0.48	0.28	1

Table 9.29: Hueng, T. T., Lee, I. H., Guog, Y. J., Chen, K. C., Chen, S. S., Chuang, S. P., ... & Yang, Y. K. (2011). Is a patient-administered depression rating scale valid for detecting cognitive deficits in patients with major depressive disorder? *Psychiatry and Clinical Neurosciences*, *65*(1), 70-76. doi:10.1111/j.1440-1819.2010.02166.x

N = 40

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	LMI	LMII
AGE	1	0.039	-0.552	-0.315	-0.279
SEX	0.039	1	0.068	-0.143	-0.13
EDU	-0.552	0.068	1	0.578	0.489
LMI	-0.315	-0.143	0.578	1	0.896
LMII	-0.279	-0.13	0.489	0.896	1

Table 9.30: Karagiannopoulou, L., Karamaouna, P., Zouraraki, C., Roussos, P., Bitsios, P., & Giakoumaki, S. G. (2016). Cognitive profiles of schizotypal dimensions in a community cohort: Common properties of differential manifestations. *Journal of Clinical and Experimental Neuropsychology*, *38*(9), 1050-1063. doi:10.1080/13803395.2016.1188890

N = 200

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LF	SF
AGE	1	0.015	0.3	0.022	0.077	0.151	0.126
SEX	0.015	1	0.264	0.042	0.009	0.218	0.091
EDU	0.3	0.264	1	-0.28	-0.314	0.415	0.409
TMTA	0.022	0.042	-0.28	1	0.582	-0.239	-0.19
TMTB	0.077	0.009	-0.314	0.582	1	-0.187	-0.207
LF	0.151	0.218	0.415	-0.239	-0.187	1	0.491
SF	0.126	0.091	0.409	-0.19	-0.207	0.491	1

Table 9.31: Kesse-Guyot, E., Andreeva, V. A., Lassale, C., Hercberg, S., & Galan, P. (2014). Clustering of midlife lifestyle behaviors and subsequent cognitive function: a longitudinal study. *American Journal of Public Health*, *104*(11), 170-177. doi:10.2105/AJPH.2014.302121

N = 2470

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LF	SF	DSF	DSB
AGE	1	-0.11	-0.1	0.25	0.24	-0.08	-0.16	-0.09	-0.06
SEX	-0.11	1	-0.04	0.01	0.02	0.12	0.04	-0.06	-0.01
EDU	-0.1	-0.04	1	-0.15	-0.28	0.29	0.27	0.18	0.2
TMTA	0.25	0.01	-0.15	1	0.49	-0.18	-0.23	-0.12	-0.15
TMTB	0.24	0.02	-0.28	0.49	1	-0.32	-0.32	-0.25	-0.31
LF	-0.08	0.12	0.29	-0.18	-0.32	1	0.5	0.26	0.27
SF	-0.16	0.04	0.27	-0.23	-0.32	0.5	1	0.22	0.24
DSF	-0.09	-0.06	0.18	-0.12	-0.25	0.26	0.22	1	0.46
DSB	-0.06	-0.01	0.2	-0.15	-0.31	0.27	0.24	0.46	1

Table 9.32: Kim, J., Jeong, J. H., Han, S. H., Ryu, H. J., Lee, J. Y., Ryu, S. H., ... & Choi, S. H. (2013). Reliability and validity of the short form of the literacy-independent cognitive assessment in the elderly. *Journal of Clinical Neurology*, *9*(2), 111-117. doi:10.3988/jcn.2013.9.2.111

N = 639

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	SF	VLT-TR	VLT-DR
AGE	1	-0.01	-0.255	-0.256	-0.337	-0.317
SEX	-0.01	1	-0.409	-0.216	0.182	0.25
EDU	-0.255	-0.409	1	0.37	0.2	0.049
SF	-0.256	-0.216	0.37	1	0.263	0.258
VLT-TR	-0.337	0.182	0.2	0.263	1	0.642
VLT-DR	-0.317	0.25	0.049	0.258	0.642	1

Table 9.33: Komulainen, P., Pedersen, M., Hänninen, T., Bruunsgaard, H., Lakka, T. A., Kivipelto, M., ... & Rauramaa, R. (2008). BDNF is a novel marker of cognitive function in ageing women: the DR's EXTRA Study. *Neurobiology of Learning and Memory*, *90*(4), 596-603. doi:10.1016/j.nlm.2008.07.014

N = 1388

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	SF	BNT	VLT-TR	VLT-DR
AGE	1	0.028	-0.194	-0.189	-0.208	-0.265	-0.24
SEX	0.028	1	0.026	-0.01	-0.23	0.219	0.168
EDU	-0.194	0.026	1	0.267	0.347	0.358	0.291
SF	-0.189	-0.01	0.267	1	0.375	0.401	0.343
BNT	-0.208	-0.23	0.347	0.375	1	0.253	0.238
VLT-TR	-0.265	0.219	0.358	0.401	0.253	1	0.746
VLT-DR	-0.24	0.168	0.291	0.343	0.238	0.746	1

Table 9.34: Krueger, K. R., Wilson, R. S., Bennett, D. A., & Aggarwal, N. T. (2009). A battery of tests for assessing cognitive function in older Latino persons. *Alzheimer Disease and Associated Disorders*, *23*(4), 384. doi:10.1097/WAD.0b013e31819e0bfc

N = 66

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	LMI	LMII	SF	DSF	DSB	COD	BNT	VLT-TR	VLT-DR
AGE	1	0.07	0.092	-0.393	-0.304	-0.346	-0.141	-0.044	-0.095	-0.037	-0.352	-0.27
SEX	0.07	1	-0.019	-0.164	-0.203	-0.047	0.097	-0.097	-0.106	0.06	-0.106	-0.088
EDU	0.092	-0.019	1	0.058	0.208	0.147	0.036	0.243	0.371	0.459	0.259	0.273
LMI	-0.393	-0.164	0.058	1	0.89	0.467	0.055	0.41	0.373	0.445	0.591	0.557
LMII	-0.304	-0.203	0.208	0.89	1	0.403	0.063	0.42	0.464	0.512	0.62	0.573
SF	-0.346	-0.047	0.147	0.467	0.403	1	0.201	0.411	0.428	0.444	0.557	0.502
DSF	-0.141	0.097	0.036	0.055	0.063	0.201	1	0.303	0.103	0.229	0.253	0.072
DSB	-0.044	-0.097	0.243	0.41	0.42	0.411	0.303	1	0.467	0.407	0.422	0.27
COD	-0.095	-0.106	0.371	0.373	0.464	0.428	0.103	0.467	1	0.579	0.424	0.386
BNT	-0.037	0.06	0.459	0.445	0.512	0.444	0.229	0.407	0.579	1	0.478	0.516
VLT-TR	-0.352	-0.106	0.259	0.591	0.62	0.557	0.253	0.422	0.424	0.478	1	0.656
VLT-DR	-0.27	-0.088	0.273	0.557	0.573	0.502	0.072	0.27	0.386	0.516	0.656	1

Table 9.35: Laukka, E. J., Lövdén, M., Herlitz, A., Karlsson, S., Ferencz, B., Pantzar, A., ... & Bäckman, L. (2013). Genetic effects on old-age cognitive functioning: a population-based study. *Psychology and Aging*, *28*(1), 262. doi:10.1037/a0030829

N = 2694

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	LF	SF
AGE	1	0.132	-0.347	-0.218	-0.452
SEX	0.132	1	-0.123	0.012	-0.016
EDU	-0.347	-0.123	1	0.352	0.355
LF	-0.218	0.012	0.352	1	0.498
SF	-0.452	-0.016	0.355	0.498	1

Table 9.36: Lehrner, J., Moser, D., Klug, S., Gleiss, A., Auff, E., Pirker, W., & Pusswald, G. (2014). Subjective memory complaints, depressive symptoms and cognition in Parkinson's disease patients. *European Journal of Neurology*, *21*(10), 1276-1285. doi:10.1111/ene.12470

N = 247

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LF	SF	COD	BNT
AGE	1	0.016	-0.213	0.309	0.367	-0.128	-0.282	-0.427	-0.169
SEX	0.016	1	-0.117	0.148	0.128	-0.162	-0.27	-0.06	-0.036
EDU	-0.213	-0.117	1	-0.177	-0.293	0.288	0.178	0.206	0.101
TMTA	0.309	0.148	-0.177	1	0.661	-0.385	-0.371	-0.566	-0.197
TMTB	0.367	0.128	-0.293	0.661	1	-0.399	-0.362	-0.599	-0.199
LF	-0.128	-0.162	0.288	-0.385	-0.399	1	0.496	0.478	0.194
SF	-0.282	-0.27	0.178	-0.371	-0.362	0.496	1	0.514	0.316
COD	-0.427	-0.06	0.206	-0.566	-0.599	0.478	0.514	1	0.233
BNT	-0.169	-0.036	0.101	-0.197	-0.199	0.194	0.316	0.233	1

Table 9.37: Liebel, S. W., Jones, E. C., Oshri, A., Hallowell, E. S., Jerskey, B. A., Gunstad, J., & Sweet, L. H. (2017). Cognitive processing speed mediates the effects of cardiovascular disease on executive functioning. *Neuropsychology*, *31*(1), 44-51. doi:10.1037/neu0000324

N = 73

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTB	LF	SF	COD
AGE	1	0.006	-0.152	0.54	-0.15	-0.45	-0.447
SEX	0.006	1	-0.159	-0.068	0.278	0.158	0.141
EDU	-0.152	-0.159	1	-0.198	0.029	0.193	0.179
TMTB	0.54	-0.068	-0.198	1	-0.366	-0.474	-0.627
LF	-0.15	0.278	0.029	-0.366	1	0.618	0.324
SF	-0.45	0.158	0.193	-0.474	0.618	1	0.465
COD	-0.447	0.141	0.179	-0.627	0.324	0.465	1

Table 9.38: Llinàs-Reglà, J., Vilalta-Franch, J., López-Pousa, S., Calvó-Perxas, L., Torrents Rodas, D., & Garre-Olmo, J. (2017). The trail making test: Association with other neuropsychological measures and normative values for adults aged 55 years and older From a Spanish-speaking population-based sample. *Assessment*, *24*(2), 183-196. doi:10.1177/1073191115602552

N = 1923

Sex coding: female > male

Education coding: lower is better

	AGE	SEX	EDU	TMTA	TMTB	LF	SF	DSF	DSB	COD
AGE	1	-0.035	0.132	0.393	0.415	-0.152	-0.235	-0.158	-0.179	-0.469
SEX	-0.035	1	0.134	0.083	0.086	-0.069	-0.069	-0.086	-0.135	-0.041
EDU	0.132	0.134	1	0.32	0.432	-0.312	-0.257	-0.303	-0.342	-0.522
TMTA	0.393	0.083	0.32	1	0.701	-0.304	-0.325	-0.286	-0.344	-0.654
TMTB	0.415	0.086	0.432	0.701	1	-0.393	-0.361	-0.379	-0.472	-0.708
LF	-0.152	-0.069	-0.312	-0.304	-0.393	1	0.466	0.316	0.368	0.422
SF	-0.235	-0.069	-0.257	-0.325	-0.361	0.466	1	0.238	0.302	0.401
DSF	-0.158	-0.086	-0.303	-0.286	-0.379	0.316	0.238	1	0.573	0.387
DSB	-0.179	-0.135	-0.342	-0.344	-0.472	0.368	0.302	0.573	1	0.458
COD	-0.469	-0.041	-0.522	-0.654	-0.708	0.422	0.401	0.387	0.458	1

Table 9.39: Mohn, C., Lystad, J. U., Ueland, T., Falkum, E., & Rund, B. R. (2017). Factor analyzing the Norwegian MATRICS consensus cognitive battery. *Psychiatry and Clinical Neurosciences*, *71*(5), 336-345. doi:10.1111/pcn.12513

N = 300

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	TMTA	SF	COD	VLT-TR
AGE	1	-0.007	0.262	0.329	0.089	-0.473	-0.289
SEX	-0.007	1	-0.176	0.002	-0.104	-0.236	-0.166
EDU	0.262	-0.176	1	0.062	0.139	0.037	0.063
TMTA	0.329	0.002	0.062	1	-0.176	-0.503	-0.225
SF	0.089	-0.104	0.139	-0.176	1	0.191	0.25
COD	-0.473	-0.236	0.037	-0.503	0.191	1	0.449
VLT-TR	-0.289	-0.166	0.063	-0.225	0.25	0.449	1

Table 9.40: Morrens, M., Hulstijn, W., Matton, C., Madani, Y., Van Bouwel, L., Peuskens, J., & Sabbe, B. G. C. (2008). Delineating psychomotor slowing from reduced processing speed in schizophrenia. *Cognitive Neuropsychiatry*, *13*(6), 457-471. doi:10.1080/13546800802439312

N = 26

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	VLT-TR	VLT-DR
AGE	1	0.182	-0.04	-0.04	0.17
SEX	0.182	1	0	-0.049	-0.024
EDU	-0.04	0	1	-0.011	-0.237
VLT-TR	-0.04	-0.049	-0.011	1	0.719
VLT-DR	0.17	-0.024	-0.237	0.719	1

Table 9.41: Ojeda, N., Pena, J., Schretlen, D. J., Sanchez, P., Aretouli, E., Elizagarate, E., ... & Gutierrez, M. (2012). Hierarchical structure of the cognitive processes in schizophrenia: the fundamental role of processing speed. *Schizophrenia Research*, *135*(1), 72-78. doi:10.1016/j.schres.2011.12.004

N = 53

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	LMI	LMII	DSB	COD
AGE	1	-0.172	-0.243	0.625	-0.04	-0.053	-0.441	-0.589
SEX	-0.172	1	0.168	0.103	0.334	0.33	0.134	0.254
EDU	-0.243	0.168	1	-0.12	0.354	0.353	0.235	0.326
TMTA	0.625	0.103	-0.12	1	-0.05	-0.044	-0.502	-0.46
LMI	-0.04	0.334	0.354	-0.05	1	0.962	0.19	0.275
LMII	-0.053	0.33	0.353	-0.044	0.962	1	0.156	0.305
DSB	-0.441	0.134	0.235	-0.502	0.19	0.156	1	0.294
COD	-0.589	0.254	0.326	-0.46	0.275	0.305	0.294	1

Table 9.42: de Paula, J. J., Bertola, L., Avila, R. T., Moreira, L., Coutinho, G., de Moraes, E. N., ... & Malloy-Diniz, L. F. (2013). Clinical applicability and cutoff values for an unstructured neuropsychological assessment protocol for older adults with low formal education. *PLoS One*, *8*(9), 1-9. doi:10.1371/journal.pone.0073167

N = 96

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	LF	SF	DSF	DSB	VLT-TR	VLT-DR
AGE	1	-0.141	-0.163	0.016	-0.085	-0.111	-0.179	-0.153	-0.018
SEX	-0.141	1	0.274	0.091	0.305	0.212	0.098	0.176	0.254
EDU	-0.163	0.274	1	0.411	0.564	0.211	0.401	0.447	0.309
LF	0.016	0.091	0.411	1	0.649	0.332	0.321	0.406	0.409
SF	-0.085	0.305	0.564	0.649	1	0.247	0.35	0.537	0.596
DSF	-0.111	0.212	0.211	0.332	0.247	1	0.246	0.34	0.256
DSB	-0.179	0.098	0.401	0.321	0.35	0.246	1	0.101	0.157
VLT-TR	-0.153	0.176	0.447	0.406	0.537	0.34	0.101	1	0.689
VLT-DR	-0.018	0.254	0.309	0.409	0.596	0.256	0.157	0.689	1

Table 9.43: Reppermund, S., Sachdev, P. S., Crawford, J., Kochan, N. A., Slavin, M. J., Kang, K., ... & Brodaty, H. (2011). The relationship of neuropsychological function to instrumental activities of daily living in mild cognitive impairment. *International Journal of Geriatric Psychiatry*, *26*(8), 843-852. doi:10.1002/gps.2612

N = 469

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LMII	LF	SF	BNT	VLT-TR	VLT-DR
AGE	1	0.037	-0.088	0.279	0.347	-0.137	-0.13	-0.255	-0.254	-0.246	-0.208
SEX	0.037	1	-0.157	0.066	0.06	0.13	-0.043	0.018	-0.079	0.258	0.229
EDU	-0.088	-0.157	1	-0.118	-0.208	0.212	0.39	0.262	0.193	0.148	-0.023
TMTA	0.279	0.066	-0.118	1	0.511	-0.039	-0.124	-0.193	-0.15	-0.117	-0.137
TMTB	0.347	0.06	-0.208	0.511	1	-0.129	-0.326	-0.274	-0.195	-0.231	-0.122
LMII	-0.137	0.13	0.212	-0.039	-0.129	1	0.1	0.228	0.167	0.388	0.314
LF	-0.13	-0.043	0.39	-0.124	-0.326	0.1	1	0.318	0.264	0.192	0.095
SF	-0.255	0.018	0.262	-0.193	-0.274	0.228	0.318	1	0.299	0.259	0.126
BNT	-0.254	-0.079	0.193	-0.15	-0.195	0.167	0.264	0.299	1	0.197	0.151
VLT-TR	-0.246	0.258	0.148	-0.117	-0.231	0.388	0.192	0.259	0.197	1	0.756
VLT-DR	-0.208	0.229	-0.023	-0.137	-0.122	0.314	0.095	0.126	0.151	0.756	1

Table 9.44: Ricarte, J. J., Ros, L., Latorre, J. M., Muñoz, M. D., Aguilar, M. J., & Hernandez, J. V. (2016). Role of anxiety and brooding in specificity of autobiographical recall. *Scandinavian Journal of Psychology*, *57*(6), 495-500. doi:10.1111/sjop.12323

N = 210

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	LF	SF
AGE	1	0.138	-0.886	-0.37	-0.545
SEX	0.138	1	0.011	0.089	0.01
EDU	-0.886	0.011	1	0.291	0.436
LF	-0.37	0.089	0.291	1	0.669
SF	-0.545	0.01	0.436	0.669	1

Table 9.45: Royall, D. R., Bishnoi, R. J., & Palmer, R. F. (2015). Serum IGF-BP2 strongly moderates age's effect on cognition: a MIMIC analysis. *Neurobiology of Aging*, *36*(7), 2232-2240. doi:10.1016/j.neurobiolaging.2015.04.003

N = 875

Sex coding: female > male

Education coding: higher is better

Table 9.46: Schmidt, C. S., Schumacher, L. V., Römer, P., Leonhart, R., Beume, L., Martin, M., ... & Kaller, C. P. (2017). Are semantic and phonological fluency based on the same or distinct sets of cognitive processes? Insights from factor analyses in healthy adults and stroke patients. *Neuropsychologia*, *99*, 148-155. doi:10.1016/j.neuropsychologia.2017.02.019

N = 69

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	LF	SF
AGE	1	-0.001	0.855	0.003	0.25
SEX	-0.001	1	-0.041	0.194	0.133
EDU	0.855	-0.041	1	-0.032	0.189
LF	0.003	0.194	-0.032	1	0.521
SF	0.25	0.133	0.189	0.521	1

Table 9.47: Siedlecki, K. L., Manly, J. J., Brickman, A. M., Schupf, N., Tang, M. X., & Stern, Y. (2010). Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers?. *Neuropsychology*, *24(3)*, 402-411. doi:10.1037/a0017515

N = 2113

Sex coding: female > male

Education coding: higher is better

Table 9.48: Snitz, B. E., Yu, L., Crane, P. K., Chang, C. C. H., Hughes, T. F., & Ganguli, M. (2012). Subjective cognitive complaints of older adults at the population level: an item response theory analysis. *Alzheimer disease and associated disorders*, *26*(4), 344-351. doi:10.1097/WAD.0b013e3182420bdf

N = 1356

Sex coding: Sex not included

Education coding: higher is better

	AGE	EDU	TMTA	TMTB	LMI	LMII	LF	SF	DSF	BNT
AGE	1	-0.189	0.357	0.417	-0.318	-0.289	-0.188	-0.325	-0.162	-0.3
EDU	-0.189	1	-0.107	-0.187	0.186	0.168	0.2	0.168	0.12	0.232
TMTA	0.357	-0.107	1	0.546	-0.098	-0.07	-0.192	-0.256	-0.154	-0.247
TMTB	0.417	-0.187	0.546	1	-0.3	-0.268	-0.28	-0.364	-0.242	-0.343
LMI	-0.318	0.186	-0.098	-0.3	1	0.872	0.246	0.364	0.225	0.391
LMII	-0.289	0.168	-0.07	-0.268	0.872	1	0.243	0.357	0.203	0.376
LF	-0.188	0.2	-0.192	-0.28	0.246	0.243	1	0.487	0.239	0.356
SF	-0.325	0.168	-0.256	-0.364	0.364	0.357	0.487	1	0.217	0.452
DSF	-0.162	0.12	-0.154	-0.242	0.225	0.203	0.239	0.217	1	0.222
BNT	-0.3	0.232	-0.247	-0.343	0.391	0.376	0.356	0.452	0.222	1

Table 9.49: Tractenberg, R. E., Fillenbaum, G., Aisen, P. S., Liebke, D. E., Yumoto, F., & Kuchibhatla, M. N. (2010). What the CERAD battery can tell us about executive function as a higher-order cognitive faculty. *Current Gerontology and Geriatrics Research*, *510614*, 1-10. doi:10.1155/2010/510614

N = 918

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	SF	BNT	VLT-TR	VLT-DR
AGE	1	-0.042	-0.428	-0.425	-0.509	-0.546	-0.533
SEX	-0.042	1	0.037	-0.024	-0.072	0.211	0.169
EDU	-0.428	0.037	1	0.518	0.584	0.567	0.496
SF	-0.425	-0.024	0.518	1	0.546	0.559	0.524
BNT	-0.509	-0.072	0.584	0.546	1	0.6	0.53
VLT-TR	-0.546	0.211	0.567	0.559	0.6	1	0.801
VLT-DR	-0.533	0.169	0.496	0.524	0.53	0.801	1

Table 9.50: Tse, C. S., Balota, D. A., Yap, M. J., Duchek, J. M., & McCabe, D. P. (2010). Effects of healthy aging and early stage dementia of the Alzheimer's type on components of response time distributions in three attention tasks. *Neuropsychology*, *24*(3), 300-315. doi:10.1037/a0018274

N = 246

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LMI	LF	SF	DSF	DSB	COD	BNT
AGE	1	-0.061	0.021	0.342	0.355	-0.157	0.044	-0.245	0.056	-0.002	-0.284	-0.007
SEX	-0.061	1	-0.159	0.008	0.019	0.015	-0.011	-0.083	-0.03	0.1	0.182	0.049
EDU	0.021	-0.159	1	-0.134	-0.205	0.247	0.284	0.199	0.104	0.063	0.119	0.102
TMTA	0.342	0.008	-0.134	1	0.673	-0.238	-0.277	-0.354	-0.045	-0.14	-0.582	-0.071
TMTB	0.355	0.019	-0.205	0.673	1	-0.251	-0.257	-0.251	-0.162	-0.271	-0.508	-0.03
LMI	-0.157	0.015	0.247	-0.238	-0.251	1	0.186	0.246	0.074	0.195	0.23	0.097
LF	0.044	-0.011	0.284	-0.277	-0.257	0.186	1	0.394	0.234	0.312	0.355	0.08
SF	-0.245	-0.083	0.199	-0.354	-0.251	0.246	0.394	1	0.19	0.217	0.344	0.213
DSF	0.056	-0.03	0.104	-0.045	-0.162	0.074	0.234	0.19	1	0.485	0.053	0.105
DSB	-0.002	0.1	0.063	-0.14	-0.271	0.195	0.312	0.217	0.485	1	0.164	0.129
COD	-0.284	0.182	0.119	-0.582	-0.508	0.23	0.355	0.344	0.053	0.164	1	0.023
BNT	-0.007	0.049	0.102	-0.071	-0.03	0.097	0.08	0.213	0.105	0.129	0.023	1

Table 9.51: Tuokko, H. A., Chou, P. H. B., Bowden, S. C., Simard, M., Ska, B., & Crossley, M. (2009). Partial measurement equivalence of French and English versions of the Canadian Study of Health and Aging neuropsychological battery. *Journal of the International Neuropsychological Society*, *15*(3), 416-425. doi:10.1017/S1355617709090602

N = 786

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	LF	SF	COD	VLT-TR
AGE	1	0.07	0.06	-0.02	-0.16	-0.28	-0.3
SEX	0.07	1	0.05	0.15	-0.06	0.07	0.25
EDU	0.06	0.05	1	0.51	0.33	0.51	0.3
LF	-0.02	0.15	0.51	1	0.48	0.58	0.42
SF	-0.16	-0.06	0.33	0.48	1	0.51	0.37
COD	-0.28	0.07	0.51	0.58	0.51	1	0.52
VLT-TR	-0.3	0.25	0.3	0.42	0.37	0.52	1

Table 9.52: Valenzuela, M. J., & Sachdev, P. (2007). Assessment of complex mental activity across the lifespan: development of the Lifetime of Experiences Questionnaire (LEQ). *Psychological Medicine*, 37(7), 1015-1025. doi:10.1017/S003329170600938X

N = 73

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LMI	LMII	LF	SF	DSF	DSB	COD	BNT
AGE	1	0.031	-0.168	0.459	0.493	-0.302	-0.362	-0.178	-0.47	0.072	-0.118	-0.482	-0.309
SEX	0.031	1	-0.3	0.139	0.017	0.078	0.011	0.227	0.07	-0.137	-0.097	-0.027	0.072
EDU	-0.168	-0.3	1	-0.185	-0.137	0.17	0.259	0.099	0.126	0.012	0.155	0.309	0.037
TMTA	0.459	0.139	-0.185	1	0.546	-0.039	-0.183	-0.056	-0.344	0.029	0.046	-0.457	-0.258
TMTB	0.493	0.017	-0.137	0.546	1	-0.289	-0.403	-0.244	-0.276	-0.142	-0.299	-0.606	-0.387
LMI	-0.302	0.078	0.17	-0.039	-0.289	1	0.895	0.154	0.231	0.097	0.325	0.39	0.425
LMII	-0.362	0.011	0.259	-0.183	-0.403	0.895	1	0.159	0.273	0.122	0.355	0.537	0.5
LF	-0.178	0.227	0.099	-0.056	-0.244	0.154	0.159	1	0.365	0.253	0.313	0.407	0.067
SF	-0.47	0.07	0.126	-0.344	-0.276	0.231	0.273	0.365	1	-0.107	0.074	0.398	0.284
DSF	0.072	-0.137	0.012	0.029	-0.142	0.097	0.122	0.253	-0.107	1	0.498	0.222	0.091
DSB	-0.118	-0.097	0.155	0.046	-0.299	0.325	0.355	0.313	0.074	0.498	1	0.249	0.221
COD	-0.482	-0.027	0.309	-0.457	-0.606	0.39	0.537	0.407	0.398	0.222	0.249	1	0.288
BNT	-0.309	0.072	0.037	-0.258	-0.387	0.425	0.5	0.067	0.284	0.091	0.221	0.288	1

Table 9.53: Waldinger, R. J., Cohen, S., Schulz, M. S., & Crowell, J. A. (2015). Security of attachment to spouses in late life: Concurrent and prospective links with cognitive and emotional well-being. *Clinical Psychological Science*, 3(4), 516-529. doi:10.1177/2167702614541261

N = 240

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	TMTA	TMTB	LF	SF	BNT	VLT-TR	VLT-DR
AGE	1	0.335	0.514	0.266	0.247	0.08	-0.269	-0.059	-0.053	-0.186
SEX	0.335	1	0.139	0.124	0.105	-0.136	-0.319	0.078	-0.121	-0.164
EDU	0.514	0.139	1	0.046	0.009	0.31	-0.008	0.033	0.227	0.064
TMTA	0.266	0.124	0.046	1	0.581	-0.336	-0.387	-0.224	-0.269	-0.293
TMTB	0.247	0.105	0.009	0.581	1	-0.338	-0.443	-0.126	-0.4	-0.408
LF	0.08	-0.136	0.31	-0.336	-0.338	1	0.553	0.108	0.443	0.318
SF	-0.269	-0.319	-0.008	-0.387	-0.443	0.553	1	0.301	0.559	0.521
BNT	-0.059	0.078	0.033	-0.224	-0.126	0.108	0.301	1	0.293	0.252
VLT-TR	-0.053	-0.121	0.227	-0.269	-0.4	0.443	0.559	0.293	1	0.74
VLT-DR	-0.186	-0.164	0.064	-0.293	-0.408	0.318	0.521	0.252	0.74	1

Table 9.54: Watts, A. S., Loskutova, N., Burns, J. M., & Johnson, D. K. (2013). Metabolic syndrome and cognitive decline in early Alzheimer's disease and healthy older adults. *Journal of Alzheimer's Disease*, *35*(2), 253-265. doi:10.3233/JAD-121168

N = 73

Sex coding: male > female

Education coding: higher is better

	AGE	SEX	EDU	LMI	LMII
AGE	1	-0.051	-0.072	-0.264	-0.296
SEX	-0.051	1	0.284	-0.138	-0.165
EDU	-0.072	0.284	1	0.153	0.081
LMI	-0.264	-0.138	0.153	1	0.847
LMII	-0.296	-0.165	0.081	0.847	1

Table 9.55: Wettstein, M., Kuźma, E., Wahl, H. W., & Heyl, V. (2016). Cross-sectional and longitudinal relationship between neuroticism and cognitive ability in advanced old age: The moderating role of severe sensory impairment. *Aging & Mental Health*, *20*(9), 918-929. doi:10.1080/13607863.2015.1049119

N = 150

Sex coding: female > male

Education coding: higher is better

	AGE	SEX	EDU	SF	DSB
AGE	1	0.077	0.124	-0.19	-0.061
SEX	0.077	1	-0.009	0.007	0.132
EDU	0.124	-0.009	1	0.262	0.124
SF	-0.19	0.007	0.262	1	0.198
DSB	-0.061	0.132	0.124	0.198	1

Table 9.56: Williams, P. G., Suchy, Y., & Kraybill, M. L. (2010). Five-factor model personality traits and executive functioning among older adults. *Journal of Research in Personality*, *44*(4), 485-491. doi:10.1016/j.jrp.2010.06.002

N = 62

Sex coding: female > male

Education coding: higher is better

10

SUPPLEMENTARY MATERIALS ACCOMPANYING CHAPTER 6: PREDICTING PARKINSON'S DISEASE DEMENTIA USING MODERN NEUROPSYCHOLOGICAL TECHNIQUES

10.1 SAMPLE CHARACTERISTICS OF THE PD GROUP FROM BROEDERS ET AL. (2013) IN TABLE 1

10.2 COGNITIVE DOMAINS

We also explored which cognitive domains were most often impaired in PD patients who were MNC-impaired, and whether there was a distinct profile for the patients who develop PDD. Figure 1 shows the mean demographically corrected z-scores at baseline. Negative z-scores indicate worse performance than the norm. From the figure, it can be observed that those who were MNC-impaired at baseline (red and blue solid lines), mainly showed impairment on the Rivermead Behavioural Memory Test and were slightly more impaired on the TMT a and the WAIS-R Digit Symbol Coding task. The WAIS-R Digit Symbol Coding task seemed to be low for all groups which is probably due to Parkinson pathology affecting motor performance. However, these tests did not discriminate very well between the PD patients who developed PDD after 5 years and those who did not. Those who were MNC-impaired and who developed PDD (red solid line) after 5 years are distinguished by low scores on the Auditory Verbal Learning subtests and Letter Fluency. These tests seem to discriminate well between those who develop PDD and those who do not. Figure 2 plots a line for every individual patient, and thus provides information on individual differences.

Table 10.1: Sample characteristics of the PD group from Broeders et al. (2013)

		N	% Men	Age range at baseline
Baseline with NPA*		123	54%	32 - 84
	attrition = 26			
3-year follow-up		97	54%	35 - 84
	attrition = 24			
5-year follow-up		73	55%	35 - 84

*NPA = Neuropsychological Assessment

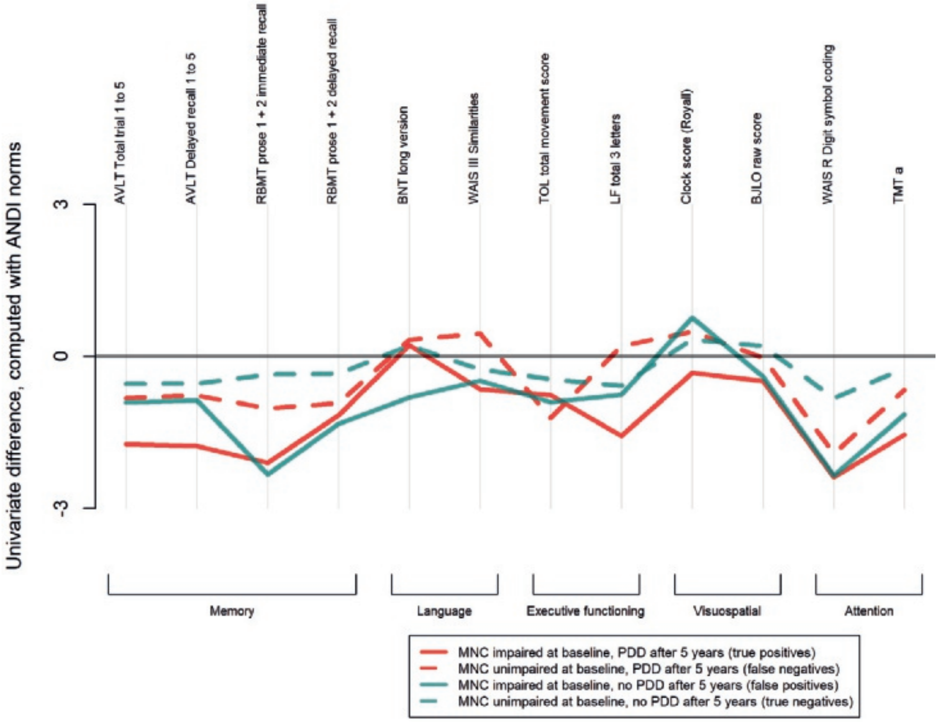

Figure 10.1: Mean demographically corrected z-scores for PD patients at baseline. Red indicates PDD after 5 years. Blue indicates no PDD after five years. The solid line is MNC-impaired, dashed is not MNC-unimpaired.

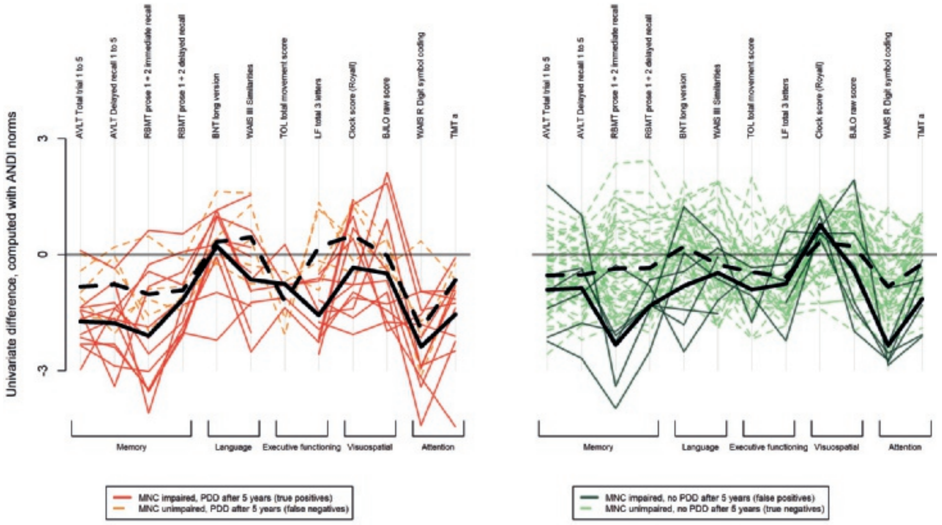

Figure 10.2: Score profiles of individual patients, in terms of differences between expected scores and observed scores. The left panel shows the patients who developed PDD after 5 years. The right panel shows those that did not develop PDD after 5 years. Solid lines: patients who are MNC impaired at baseline, dashed lines: patients who are not MNC impaired at baseline. The black lines denote the matching mean scores.
Note that not all patients completed all tests. Therefore, some lines are interrupted.

Table 10.2: A comparison of the classifications between the traditional PD-MCI criteria and the MNC method applied with ANDI. Note that the number of PDD cases and missing cases are cumulative.

	ANDI MNC impaired			ANDI MNC not impaired		
		3 years	5 years		3 years	5 years
PD-MCI	24	6 PDD	8 PDD	19	1 PDD	2 PDD
		12 no PDD	4 no PDD		16 no PDD	10 no PDD
		6 missing	12 missing		2 missing	7 missing
no PD-MCI	8	2 PDD	4 PDD	72	0 PDD	3 PDD
		4 no PDD	2 no PDD		56 no PDD	40 no PDD
		2 missing	2 missing		16 missing	29 missing

10.3 overlap in diagnosis between methods

We investigated whether the patients who were diagnosed as impaired were the same across methods, or whether there were differences. We also examined how differences in who was diagnosed as impaired, can explain the differences in how well methods perform in the prediction of progression to PDD after three and five years. As can be seen in Table 2, there were 19 PD patients who were unimpaired according to the MNC method applied with ANDI, whereas they were impaired according to the original PD-MCI method. One of these patients progressed to PDD after three years, and two progressed to PDD after five years. Although the number of patients impaired according to the PD-MCI method (N=43) was higher than the number impaired according to the MNC method (N=32), there were still eight patients who were only diagnosed as impaired by the MNC method. Of these eight patients, two progressed to PDD after three years, and four progressed to PDD after five years. Therefore, these eight patients seem to be an important subgroup that was missed with the traditional method. Overall, there was a moderate degree of agreement between methods (78%, κ = 0.49).

We made a similar comparison of the two methods that make use of the ANDI database. The PD-MCI criteria applied with ANDI and the MNC methods applied with ANDI yielded different results, although there was a good degree of agreement between methods (86%, κ = 0.63). As can be seen in Table 3, nine patients had PD-MCI according to the criteria applied with ANDI but are not abnormal according to the MNC method. None of these patients developed dementia after three or five years. Eight patients were MNC-abnormal but did not have PD-MCI according to the criteria. Of these eight, two developed PDD after three years, and two more patients (four in total) had de-

Table 10.3: A comparison of the classifications between the PD-MCI criteria applied with ANDI and the MNC method applied with ANDI. Note that the number of PDD cases and missing cases are cumulative.

	ANDI MNC impaired			ANDI MNC not impaired		
		3 years	5 years		3 years	5 years
ANDI PD-MCI	24	6 PDD 12 no PDD 6 missing	8 PDD 4 no PDD 12 missing	9	0 PDD 8 no PDD 1 missing	0 PDD 4 no PDD 5 missing
ANDI no PD-MCI	8	2 PDD 4 no PDD 2 missing	4 PDD 2 no PDD 2 missing	82	1 PDD 64 no PDD 17 missing	5 PDD 46 no PDD 31 missing

veloped dementia after five years. Again, the MNC method identified some patients who would develop PDD but who were not detected by the PD-MCI method.

Last, we compared the two applications of the PD-MCI criteria. More patients were diagnosed with the traditional PD-MCI criteria than with the ANDI-MCI-criteria. There was a good degree of agreement between methods (85%, κ = 0.68). This could suggest that the ANDI PD-MCI criteria method diagnosed the same patients as the original PD-MCI criteria, but fewer. The results in Table 4 indicate that this indeed was the case to some extent. There were 13 PD patients who were unimpaired according to the PD-MCI criteria applied with ANDI, but were impaired according to the PD-MCI criteria as applied by Broeders et al. (2013). Three of these patients progressed to PDD after 5 years. The fact that the PD-MCI method with ANDI diagnosed fewer patients implies that future PDD patients were missed at baseline. However, there were also three patients who were diagnosed as PD-MCI by the PD-MCI criteria applied with ANDI who were not impaired when using the PD-MCI criteria as applied by Broeders et al. (2013). Of these three, one became demented after three years. Thus, using ANDI with the PD-MCI criteria also identified one patient who developed PDD who was missed by the traditional method.

Table 10.4: A comparison of the classifications between the traditional PD-MCI criteria and the PD-MCI criteria applied with ANDI. Note that the number of PDD cases and missing cases are cumulative.

	ANDI PD-MCI			ANDI no PD-MCI		
		3 years	5 years		3 years	5 years
PD-MCI	30	5 PDD	7 PDD	13	2 PDD	3 PDD
		20 no PDD	8 no PDD		8 no PDD	6 PDD
		5 missing	15 missing		3 missing	4 missing
no PD-MCI	3	1 PDD	1 PDD	77	1 PDD	6 PDD
		0 no PDD	0 no PDD		60 no PDD	42 no PDD
		2 missing	2 missing		16 missing	29 missing

11

REFERENCES

Aarsland, D., Andersen, K., Larsen, J. P., Lolk, A., Nielsen, H., & Kragh–Sørensen, P. (2001). Risk of dementia in Parkinson's disease: A community-based, prospective study. *Neurology*, *56*(6), 730-736.

Aarsland, D., Brønnick, K., Larsen, J. P., Tysnes, O. B., Alves, G., & Norwegian ParkWest Study Group. (2009). Cognitive impairment in incident, untreated Parkinson disease: The Norwegian ParkWest Study. *Neurology*, *72*(13), 1121-1126.

Adrover-Roig, D., Sesé, A., Barceló, F., & Palmer, A. (2012). A latent variable approach to executive control in healthy ageing. *Brain and Cognition*, *78*(3), 284-299.

Advanced Neuropsychological Diagnostics Infrastructure. (2016, August 2). Retrieved from http://www.andi.nl

Agelink van Rentergem, J. A., Murre, J. M. J., & Huizenga, H. M. (2017). Multivariate normative comparisons using an aggregated database. *PLoS ONE*, *12*, 1-18.

Agelink van Rentergem, J. A., de Vent, N. R., Schmand, B. A., Murre, J. M. J., & Huizenga, H. M. (2017). Multivariate normative comparisons for neuropsychological assessment by a multilevel factor structure or multiple imputation approach, *Psychological Assessment*. Advance online publication. doi:10.1037/pas0000489

Agresti, A., & Coull B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician, 52*, 119-126.

Albert, M., Massaro, J., DeCarli, C., Beiser, A., Seshadri, S., Wolf, P. A., & Au, R. (2010). Profiles by sex of brain MRI and cognitive function in the framingham offspring study. *Alzheimer Disease and Associated Disorders*, *24*(2), 190-193.

Andrejeva, N., Knebel, M., Dos Santos, V., Schmidt, J., Herold, C. J., Tudoran, R., ... & Gorenc-Mahmutaj, L. (2016). Neurocognitive deficits and effects of cognitive reserve in mild cognitive impairment. *Dementia and Geriatric Cognitive Disorders*, *41*(3-4), 199-209.

Andreotti, C., & Hawkins, K. A. (2015). RBANS norms based on the relationship of age, gender, education, and WRAT-3 reading to performance within an older African American sample. *The Clinical Neuropsychologist*, *29*(4), 442-465.

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optic*s, *34*(5), 502-508.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108-119.

Barnes, L. L., Yumoto, F., Capuano, A., Wilson, R. S., Bennett, D. A., & Tractenberg, R. E. (2016). Examination of the factor structure of a global cognitive function battery across race and time. *Journal of the International Neuropsychological Society*, *22*(1), 66-75.

Bartram, D. (2008). Global norms: Towards some guidelines for aggregating personality norms across countries. *International Journal of Testing*, *8*(4), 315-333.

Bennett, I. J., & Stark, C. E. (2016). Mnemonic discrimination relates to perforant path integrity: an ultra-high resolution diffusion tensor imaging study. *Neurobiology of Learning and Memory*, *129*, 107-112.

Benton, A.L. & Hamsher, K. (1983). *Multilingual Aphasia Examination.* Iowa City: AJA Associates.

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

Bezdicek, O., Libon, D. J., Stepankova, H., Panenkova, E., Lukavsky, J., Garrett, K. D., ... & Kopecek, M. (2014). Development, validity, and normative data study for the 12-word Philadelphia Verbal Learning Test [czP (r) VLT-12] among older and very old Czech adults. *The Clinical Neuropsychologist*, *28*(7), 1162-1181.

Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "Abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology, 24*(1), 31-46.

Bird, C. M., Castelli, F., Malik, O., Frith, U., & Husain, M. (2004). The impact of extensive medial frontal lobe damage on 'Theory of Mind' and cognition. *Brain*, *127*, 914-928.

Blakesley, R. E., Mazumdar, S., Dew, M. A., Houck, P. R., Tang, G., Reynolds III, C. F., & Butters, M. A. (2009). Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology, 23*(2), 255-264.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *British Medical Journal, 310*(6973), 170.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*, 127-135.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605-634.

Booth, T., Royle, N. A., Corley, J., Gow, A. J., Hernández, M. D. C. V., Maniega, S. M., ... & Deary, I. J. (2015). Association of allostatic load with brain structure and cognitive ability in later life. *Neurobiology of Aging*, *36*(3), 1390-1399.

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, *64*(9), 1089-1108.

Bouazzaoui, B., Fay, S., Taconnat, L., Angel, L., Vanneste, S., & Isingrini, M. (2013). Differential involvement of knowledge representation and executive control in episodic memory performance in young and older adults. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *67*(2), 100-107.

Bowden, S. C., Cook, M. J., Bardenhagen, F. J., Shores, E. A., & Carstairs, J. R. (2004). Measurement invariance of core cognitive abilities in heterogeneous neurological and community samples. *Intelligence*, *32*(4), 363-389.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211-252.

Broeders, M., de Bie, R. M. A., Velseboer, D. C., Speelman, J. D., Muslimovic, D., & Schmand, B. (2013). Evolution of mild cognitive impairment in Parkinson disease. *Neurology*, *81*(4), 346-352.

Broeders, M., Velseboer, D. C., de Bie, R., Speelman, J. D., Muslimovic, D., Post, B., ..., Schmand, B. (2013). Cognitive change in newly-diagnosed patients with Parkinson's disease: A 5-year follow-up study. *Journal of the International Neuropsychological Society*, *19*, 695-708.

Brooks, B. L., Iverson, G. L., & White, T. (2009). Advanced interpretation of the neuropsychological assessment battery with older adults: base rate analyses, discrepancy scores, and interpreting change. *Archives of Clinical Neuropsychology*, *24*, 647-657.

Bunce, D., Batterham, P. J., Christensen, H., & Mackinnon, A. J. (2014). Causal associations between depression symptoms and cognition in a community-based cohort of older adults. *The American Journal of Geriatric Psychiatry*, *22*(12), 1583-1591.

Burns, N. R., Nettelbeck, T., & McPherson, J. (2009). Attention and intelligence: A factor analytic study. *Journal of Individual Differences*, *30*(1), 44-57.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*.

van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox, & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173–196). New York: Routledge.

Cao, J., & Zhang, S. (2014). Multiple comparison procedures. *Journal of the American Medical Association*, *312*(5), 543-544.

Cappelletti, M., Butterworth, B., & Kopelman, M. (2012). Numeracy skills in patients with degenerative disorders and focal brain lesions: A neuropsychological investigation. *Neuropsychology*, *26*, 1-19.

Castelli, L., Rizzi, L., Zibetti, M., Angrisano, S., Lanotte, M., & Lopiano, L. (2010). Neuropsychological changes 1-year after subthalamic DBS in PD patients: A prospective controlled study. *Parkinsonism & Related Disorders*, *16*, 115-118.

Caviness, J. N., Driver-Dunckley, E., Connor, D. J., Sabbagh, M. N., Hentz, J. G., Noble, B., ..., Adler, C. H. (2007). Defining mild cognitive impairment in Parkinson's disease. *Movement Disorders*, *22*(9), 1272-1277.

Chan, R. C., Wang, Y., Wang, L., Chen, E. Y., Manschreck, T. C., Li, Z. J., ... & Gong, Q. Y. (2009). Neurological soft signs and their relationships to neurocognitive functions: A re-visit with the structural equation modeling design. *PLoS One*, *4*(12), 1-8.

Chen, Y. C., Jung, C. C., Chen, J. H., Chiou, J. M., Chen, T. F., Chen, Y. F., ... & Lee, M. S. (2017). Association of dietary patterns with global and domain-specific cognitive decline in Chinese elderly. *Journal of the American Geriatrics Society*, *65*(6), 1159-1167.

Cheng, G., Huang, C., Deng, H., & Wang, H. (2012). Diabetes as a risk factor for dementia and mild cognitive impairment: A meta-analysis of longitudinal studies. *Internal Medicine Journal*, *42*(5), 484-491.

Cheung, M. W. L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods*, *10*(1), 40-64.

Cheung, M. W. L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*(1521), 1-7.

Ciccarelli, N., Fabbiani, M., Baldonero, E., Fanti, I., Cauda, R., Giambenedetto, S. D., & Silveri, M. C. (2012). Effect of aging and human immunodeficiency virus infection on cognitive abilities. *Journal of the American Geriatrics Society*, *60*(11), 2048-2055.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, S., ter Stege, J. A., Geurtsen, G. J., Scherpbier, H. J., Kuijpers, T. W., Reiss, P., ..., Pajkrt, D. (2014). Poorer cognitive performance in perinatally HIV-infected children as compared to healthy socioeconomically matched controls. *Clinical Infectious Diseases*, *60*(7), 1111–1119.

Crawford, J. R., & Allan, K. M. (1994). The Mahalanobis Distance index of WAIS-R subtest scatter: Psychometric properties in a healthy UK sample. *British Journal of Clinical Psychology*, *33*, 65-69.

Crawford, J. R., & Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia*, *40*, 1196-1208.

Crawford, J. R., Garthwaite, P. H., Azzalini, A., Howell, D. C., & Laws, K. R. (2006). Testing for a deficit in single-case studies: effects of departures from normality. *Neuropsychologia*, *44*, 666–677.

Crawford, J. R., & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist, 12*(4), 482-486.

Cudeck, R. (2000). An estimate of the covariance between variables which are not jointly observed. *Psychometrika*, *65*, 539-546.

Culbertson, W. C., & Zillmer, E. A. (1998). The Tower of London DX: A standardized approach to assessing executive functioning in children. *Archives of Clinical Neuropsychology*, *13*(3), 285-301.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81-100.

Darst, B. F., Koscik, R. L., Hermann, B. P., La Rue, A., Sager, M. A., Johnson, S. C., & Engelman, C. D. (2015). Heritability of cognitive traits among siblings with a parental history of Alzheimer's disease. *Journal of Alzheimer's Disease*, *45*(4), 1149-1155.

Delandshere, G. (2001). Implicit theories, unexamined assumptions and the status quo of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *8*(2), 113-133.

Delis, D. C., Jacobson, M., Bondi, M. W., Hamilton, J. M., & Salmon, D. P. (2003). The myth of testing construct validity using factor analysis or correlations with normal or mixed clinical populations: Lessons from memory assessment. *Journal of the International Neuropsychological Society*, *9*(6), 936-946.

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2005). Sources of openness/intellect: Cognitive and neuropsychological correlates of the fifth factor of personality. *Journal of Personality*, *73*(4), 825-858.

Dowling, N. M., Hermann, B., La Rue, A., & Sager, M. A. (2010). Latent structure and factorial invariance of a neuropsychological test battery for the study of preclinical Alzheimer's disease. *Neuropsychology*, *24*, 742-756.

Domellöf, M. E., Ekman, U., Forsgren, L., & Elgh, E. (2015). Cognitive function in the early phase of Parkinson's disease, a five-year follow-up. *Acta Neurologica Scandinavica*, *132*(2), 79-88.

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends In Cognitive Sciences*, *20*(6), 425-443.

Duff, K. D., Langbehn, D. R., Schoenberg, M. R., Moser, D. J., Baade, L. E., Mold, J. W., ... Adams, R. L. (2006). Examining the repeatable battery for the assessment of neuropsychological status: Factor analytic studies in an elderly sample. *The American Journal of Geriatric Psychiatry*, *14*, 976-979.

Dupont, W. D., & Plummer, W. D. (1990). Power and sample size calculations: A review and

computer program. *Controlled Clinical Trials, 11*(2), 116-128.

Eifler, S., Rausch, F., Schirmbeck, F., Veckenstedt, R., Englisch, S., Meyer-Lindenberg, A., ... & Zink, M. (2014). Neurocognitive capabilities modulate the integration of evidence in schizophrenia. *Psychiatry Research*, *219*(1), 72-78.

Elgh, E., Domellöf, M., Linder, J., Edström, M., Stenlund, H., & Forsgren, L. (2009). Cognitive function in early Parkinson's disease: a population-based study. *European Journal of Neurology*, *16*(12), 1278-1284.

Emre, M., Aarsland, D., Brown, R., Burn, D. J., Duyckaerts, C., Mizuno, Y., ..., Goldman, J. (2007). Clinical diagnostic criteria for dementia associated with Parkinson's disease. *Movement Disorders*, *22*(12), 1689-1707.

Enders, C., & Bandalos, D. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 430–457.

Enders, C. K. (2006). A primer on the use of modern missing-data methods in psychosomatic medicine research. *Psychosomatic Medicine*, *68*, 427-436.

Evans, S. J., Elliott, G., Reynders, H., & Isaac, C. L. (2014). Can temporal lobe epilepsy surgery ameliorate accelerated long-term forgetting? *Neuropsychologia*, *53*, 64-74.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299.

Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology, 2*(8), 1-4.

Fernaeus, S. E., Östberg, P., Wahlund, L. O., & Hellström, Å. (2014). Memory factors in Rey AVLT: implications for early staging of cognitive decline. *Scandinavian Journal of Psychology*, *55*(6), 546-553.

Ferreira, N. V., Cunha, P. J., da Costa, D. I., dos Santos, F., Costa, F. O., Consolim-Colombo, F., & Irigoyen, M. C. (2015). Association between functional performance and executive cognitive functions in an elderly population including patients with low ankle–brachial index. *Clinical Interventions in Aging*, *10*, 839-847.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: SAGE publications.

Folstein, M.F., Folstein, S.E., McHugh, P.R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12(3)*, 189-198.

Fortin, A., & Caza, N. (2014). A validation study of memory and executive functions indexes in French-speaking healthy young and older adults. *Canadian Journal on Aging/La Revue canadienne du vieillissement*, *33*(1), 60-71.

Gallagher, P., Gray, J. M., Watson, S., Young, A. H., & Ferrier, I. N. (2014). Neurocognitive functioning in bipolar depression: a component structure analysis. *Psychological Medicine*, *44*(5), 961-974.

Galtier, I., Nieto, A., Lorenzo, J. N., & Barroso, J. (2016). Mild cognitive impairment in Parkinson's disease: Diagnosis and progression to dementia. *Journal of Clinical and Experimental Neuropsychology*, *38*(1), 40-50.

Ganguli, M., Chang, C. C. H., Snitz, B. E., Saxton, J. A., Vanderbilt, J., & Lee, C. W. (2010). Prevalence of mild cognitive impairment by multiple classifications: the Monongahela-Youghiogheny Healthy Aging Team (MYHAT) project. *The American Journal of Geriatric Psychiatry*, *18*(8), 674-683.

Gasca-Salas, C., Estanga, A., Clavero, P., Aguilar-Palacio, I., González-Redondo, R., Obeso, J. A., & Rodríguez-Oroz, M. C. (2014). Longitudinal assessment of the pattern of cognitive decline in non-demented patients with advanced Parkinson's disease. *Journal of Parkinson's Disease*, *4*(4), 677-686.

Gavett, B. E. (2015). The value of Bayes' theorem for interpreting abnormal test scores in cognitively healthy and clinical samples. *Journal of the International Neuropsychological Society*, *21*(3), 249-257.

Gordon, A. Y. (2011). A new optimality property of the Holm step-down procedure. *Statistical Methodology, 8*(2), 129-135.

Gordon, A. Y., & Salzman, P. (2008). Optimality of the Holm procedure among general step-down multiple testing procedures. *Statistics & Probability Letters, 78*(13), 1878-1884.

González-Redondo, R., Toledo, J., Clavero, P., Lamet, I., García-García, D., García-Eulate, R., ..., Rodríguez-Oroz, M. C. (2012). The impact of silent vascular brain burden in cognitive impairment in Parkinson's disease. *European Journal of Neurology*, *19*, 1100–1107

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*, 80-100.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*, 323-343.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576.

Grasman, R. P. P. P., Huizenga, H. M., & Geurts, H. M. (2010). Departure from normality in multivariate normative comparison: The Cramér alternative for Hotelling's T^2. *Neuropsychologia*, *48*, 1510–1516.

Greenaway, M. C., Smith, G. E., Tangalos, E. G., Geda, Y. E., & Ivnik, R. J. (2009). Mayo Older Americans Normative Studies: Factor analysis of an expanded neuropsychological battery. *The Clinical Neuropsychologist*, *23*, 7-20.

Gross, A. L., Mungas, D. M., Crane, P. K., Gibbons, L. E., MacKay-Brandt, A., Manly, J. J., ... & Potter, G. G. (2015). Effects of education and race on cognitive decline: An integrative study of generalizability versus study-specific results. *Psychology and Aging*, *30*(4), 863-880.

Hallquist, M., & Wiley, J. (2013). MplusAutomation: Automating Mplus model estimation and interpretation (Version 0.6-2).

Harvey, P. D. (2012). Clinical applications of neuropsychological assessment. *Dialogues in Clinical Neuroscience, 14*(1), 91-99.

Hedden, T., & Yoon, C. (2006). Individual differences in executive processing predict susceptibility to interference in verbal working memory. *Neuropsychology*, *20*(5), 511-528.

Hedden, T., Mormino, E. C., Amariglio, R. E., Younger, A. P., Schultz, A. P., Becker, J. A., ... & Rentz, D. M. (2012). Cognitive profile of amyloid burden and white matter hyperintensities in cognitively normal older adults. *Journal of Neuroscience*, *32*(46), 16233-16242.

Hobson, P., & Meara, J. (2004). The risk and incidence of dementia in a cohort of older subjects with Parkinson's disease int the UK. *Movement Disorders, 19(9)*, 1043-1049.

Hobson, P., & Meara, J. (2015). Mild cognitive impairment in Parkinson's disease and its progression onto dementia: a 16-year outcome evaluation of the Denbighshire cohort. *International Journal of Geriatric Psychiatry*, *30*(10), 1048-1055.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Hoogland, J., Boel, J. A., de Bie, R. M., Geskus, R.B., Schmand, B. A., Dalrymple-Alford, J. C., . . . , Geurtsen, G.J. (2017). Mild cognitive impairment as a risk factor for Parkinson's disease dementia. *Movement Disorders*, *32(7)*, 1056-1065.

Horvat, P., Richards, M., Malyutina, S., Pajak, A., Kubinova, R., Tamosiunas, A., ... & Bobak, M. (2014). Life course socioeconomic position and mid-late life cognitive function in Eastern Europe. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *69*(3), 470-481.

Huba, G. J. (1985). How unusual is a profile of test scores? *Journal of Psychoeducational Assessment*, *3*, 321-325.

Hueng, T. T., Lee, I. H., Guog, Y. J., Chen, K. C., Chen, S. S., Chuang, S. P., ... & Yang, Y. K. (2011). Is a patient-administered depression rating scale valid for detecting cognitive deficits in patients with major depressive disorder? *Psychiatry and Clinical Neurosciences*, *65*(1), 70-76.

Hughes, T. A., Ross, H. F., Musa, S., Bhattacherjee, S., Nathan, R. N., Mindham, R. H. S., & Spokes, E. G. S. (2000). A 10-year study of the incidence of and factors predicting dementia in Parkinson's disease. *Neurology*, *54*(8), 1596-1603.

Huizenga, H. M., Smeding, H., Grasman, R. P. P. P., & Schmand, B. (2007). Multivariate normative comparisons, *Neuropsychologia*, *45*, 2534–2542.

Huizenga, H. M., van der Molen, M. W., Bexkens, A., Bos, M. G., & van den Wildenberg, W. P. (2012). Muscle or motivation? A stop-signal study on the effects of sequential cognitive control. *Frontiers in Psychology*, *3*(126), 1-10.

Huizenga, H. M., Agelink van Rentergem, J. A., Grasman, R. P. P. P., Muslimovic, D., & Schmand, B. (2016). Normative comparisons for large neuropsychological test batteries: User-friendly and sensitive solutions to minimize familywise false positives. *Journal of Clinical and Experimental Neuropsychology*, *38*, 611-629.

Huizinga, M., Dolan, C. V., & van der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia*, *44*(11), 2017-2036.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J. M., & Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis*, *51*, 5142-5154.

Jak, S. (2015). *Meta-analytic structural equation modelling*. Springer International Publishing. doi:10.1007/978-3-319-27174-3

Janvin, C., Aarsland, D., Larsen, J. P., & Hugdahl, K. (2003). Neuropsychological profile of patients with Parkinson's disease without dementia. *Dementia and Geriatric Cognitive Disorders*, *15*(3), 126-131.

Jewsbury, P. A., Bowden, S. C., & Duff, K. (2016). The Cattell-Horn-Carroll model of cognition for clinical assessment. *Journal of Psychoeducational Assessment*, *35*(6), 547-567.

Jewsbury, P. A., & Bowden, S. C. (2016). Construct validity of fluency and implications for the factorial structure of memory. *Journal of Psychoeducational Assessment*, *35*(5), 460–481.

Kafadar, H. (2012). Cognitive model of problem solving. *New Symposium*, *50*(4), 195-206.

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

Karagiannopoulou, L., Karamaouna, P., Zouraraki, C., Roussos, P., Bitsios, P., & Giakoumaki, S. G. (2016). Cognitive profiles of schizotypal dimensions in a community cohort: Common properties of differential manifestations. *Journal of Clinical and Experimental Neuropsychology*, *38*(9), 1050-1063.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67*(3), 285-299.

Kesse-Guyot, E., Andreeva, V. A., Lassale, C., Hercberg, S., & Galan, P. (2014). Clustering of midlife lifestyle behaviors and subsequent cognitive function: a longitudinal study. *American Journal of Public Health*, *104*(11), 170-177.

Kim, J., Jeong, J. H., Han, S. H., Ryu, H. J., Lee, J. Y., Ryu, S. H., ... & Choi, S. H. (2013). Reliability and validity of the short form of the literacy-independent cognitive assessment in the elderly. *Journal of Clinical Neurology*, *9*(2), 111-117.

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science, 331*(6018), 719-721.

Kolenikov, S., & Bollen, K. A. (2012). Testing negative error variances: Is a Heywood case a symptom of misspecification? *Sociological Methods & Research*, *41*(1), 124-167.

Komulainen, P., Pedersen, M., Hänninen, T., Bruunsgaard, H., Lakka, T. A., Kivipelto, M., ... & Rauramaa, R. (2008). BDNF is a novel marker of cognitive function in ageing women: the DR's EXTRA Study. *Neurobiology of Learning and Memory*, *90*(4), 596-603.

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*, 1524-1529.

Krueger, K. R., Wilson, R. S., Bennett, D. A., & Aggarwal, N. T. (2009). A battery of tests for assessing cognitive function in older Latino persons. *Alzheimer Disease and Associated Disorders*, *23*(4), 384-388.

Larrabee, G. J. (2003). Lessons on measuring construct validity: A commentary on Delis, Jacobson, Bondi, Hamilton, and Salmon. *Journal of the International Neuropsychological Society*, *9*(6), 947-953.

Larrabee, G. J. (2014). Test validity and performance validity: Considerations in providing a framework for development of an ability-focused neuropsychological test battery. *Archives of Clinical Neuropsychology*, *29*(7), 695-714.

Laukka, E. J., Lövdén, M., Herlitz, A., Karlsson, S., Ferencz, B., Pantzar, A., ... & Bäckman, L. (2013). Genetic effects on old-age cognitive functioning: A population-based study. *Psychology and Aging*, *28*(1), 262-274.

Lehrner, J., Moser, D., Klug, S., Gleiss, A., Auff, E., Pirker, W., & Pusswald, G. (2014). Subjective memory complaints, depressive symptoms and cognition in Parkinson's disease patients. *European Journal of Neurology*, *21*(10), 1276-1285.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*, 764–766

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

Li, D., & Dye, T. D. (2013). Power and stability properties of resampling-based multiple testing procedures with applications to gene oncology studies. *Computational and Mathematical Methods in Medicine*, *610297*, 1-11.

Li, S. C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science*, *15*, 155-163.

Libon, D. J., Xie, S. X., Eppig, J., Wicas, G., Lamar, M., Lippa, C., ... & Wambach, D. M. (2010). The heterogeneity of mild cognitive impairment: A neuropsychological analysis. *Journal of the International Neuropsychological Society*, *16*(1), 84-93.

Liebel, S. W., Jones, E. C., Oshri, A., Hallowell, E. S., Jerskey, B. A., Gunstad, J., & Sweet, L. H. (2017). Cognitive processing speed mediates the effects of cardiovascular disease on executive functioning. *Neuropsychology*, *31*(1), 44-51.

Litvan, I., Goldman, J. G., Tröster, A. I., Schmand, B. A., Weintraub, D., Petersen, R. C., ..., Emre, M. (2012). Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines. *Movement Disorders*, *27*(3), 349-356.

Litvan, I., Aarsland, D., Adler, C. H., Goldman, J. G., Kulisevsky, J., Mollenhauer, B., ..., Weintraub, D. (2011). MDS task force on mild cognitive impairment in Parkinson's disease: Critical review of PD-MCI. *Movement Disorders*, *26*(10), 1814-1824.

Llinàs-Reglà, J., Vilalta-Franch, J., López-Pousa, S., Calvó-Perxas, L., Torrents Rodas, D., & Garre-Olmo, J. (2017). The trail making test: Association with other neuropsychological measures and normative values for adults aged 55 years and older from a Spanish-speaking population-based sample. *Assessment*, *24*(2), 183-196.

Maas, A. I., Stocchetti, N., & Bullock, R. (2008). Moderate and severe traumatic brain injury in adults. *The Lancet Neurology*, *7*(8), 728-741.

van der Maas, H. L., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842-861.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84-99.

Mamikonyan, E., Moberg, P. J., Siderowf, A., Duda, J. E., Ten Have, T., Hurtig, H. I., Stern, M.B,, & Weintraub, D. (2009). Mild cognitive impairment is common in Parkinson's disease patients with normal Mini-Mental State Examination (MMSE) scores. *Parkinsonism & Related Disorders*, *15*(3), 226-231.

McCaffrey, R. J., & Westervelt, H. J. (1995). Issues associated with repeated neuropsychological assessments. *Neuropsychology Review*, *5*(3), 203-221.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*, 1-10.

Meyer, A. C., Boscardin, W. J., Kwasa, J. K., & Price, R. W. (2013). Is it time to rethink how neuropsychological tests are used to diagnose mild forms of HIV-associated neurocognitive disorders? Impact of false-positive rates on prevalence and power. *Neuroepidemiology*, *41*, 208-216.

Miller, J. B., & Barr, W. B. (2017). The technology crisis in neuropsychology. *Archives of Clinical Neuropsychology*, *32*, 541-554.

Mitchell, M. B., Shaughnessy, L. W., Shirk, S. D., Yang, F. M., & Atri, A. (2012). Neuropsychological test performance and cognitive reserve in healthy aging and the Alzheimer's disease spectrum: A theoretically driven factor analysis. *Journal of the International Neuropsychological Society*, *18*, 1071-1080.

Mohn, C., Lystad, J. U., Ueland, T., Falkum, E., & Rund, B. R. (2017). Factor analyzing the Norwegian MATRICS consensus cognitive battery. *Psychiatry and Clinical Neurosciences*, *71*(5), 336-345.

Moore, A. R., & O'Keeffe, S. T. (1999). Drug-induced cognitive impairment in the elderly. *Drugs & Aging*, *15*(1), 15-28.

Moran, M. D. (2003). Arguments for rejecting the sequential Bonferroni in ecological

studies. *Oikos, 100*(2), 403-405.

Morrens, M., Hulstijn, W., Matton, C., Madani, Y., Van Bouwel, L., Peuskens, J., & Sabbe, B. G. C. (2008). Delineating psychomotor slowing from reduced processing speed in schizophrenia. *Cognitive Neuropsychiatry*, *13*(6), 457-471.

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*, *43*(11), 2412-2414.

Morrison, S. L., & Morris, J. N. (1959). Epidemiological observations on high blood-pressure without evident cause. *The Lancet*, *274*(7108), 864-870.

Muslimović, D., Post, B., Speelman, J. D., & Schmand, B. (2005). Cognitive profile of patients with newly diagnosed Parkinson disease. *Neurology*, *65*(8), 1239-1245.

Muslimović, D., Post, B., Speelman, J. D., De Haan, R. J., & Schmand, B. (2009). Cognitive decline in Parkinson's disease: a prospective longitudinal study. *Journal of the International Neuropsychological Society*, *15*(3), 426-437.

Muthén, B. (1997). Latent variable modeling with longitudinal and multilevel data. In A. Raftery (Eds.), *Sociological Methodology* (pp. 453-480). Boston: Blackwell Publishers.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide (7th ed.)*. Los Angeles: Muthén & Muthén.

Narum, S. R. (2006). Beyond Bonferroni: Less conservative analyses for conservation genetics. *Conservation Genetics, 7*(5), 783-787.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping, 15*(1), 1-25.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422-1425.

O'Carroll, R. E., Ebmeier, K. P., Dougall, N., Murray, C., Goodwin, G. M., Hayes, P. C., ... & Best, J. J. K. (1991). Regional cerebral blood flow and cognitive function in patients with chronic liver disease. *The Lancet*, *337*(8752), 1250-1253.

Ojeda, N., Pena, J., Schretlen, D. J., Sanchez, P., Aretouli, E., Elizagarate, E., ... & Gutierrez, M. (2012). Hierarchical structure of the cognitive processes in schizophrenia: the fundamental role of processing speed. *Schizophrenia Research*, *135*(1), 72-78.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(210), 1-10.

Park, L. Q., Gross, A. L., McLaren, D. G., Pa, J., Johnson, J. K., Mitchell, M., ... & Alzheimer's Disease Neuroimaging Initiative. (2012). Confirmatory factor analysis of the ADNI neuropsychological battery. *Brain Imaging and Behavior*, *6*(4), 528-539.

Parikh, R., Mathai, A., Parikh, S., Sekhar, G. C., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology, 56*(1), 45-50.

de Paula, J. J., Bertola, L., Avila, R. T., Moreira, L., Coutinho, G., de Moraes, E. N., ... & Malloy-Diniz, L. F. (2013). Clinical applicability and cutoff values for an unstructured neuropsychological assessment protocol for older adults with low formal education. *PLoS One*, *8*(9), 1-9.

Pedersen, K. F., Larsen, J. P., Tysnes, O. B., & Alves, G. (2013). Prognosis of mild cognitive impairment in early Parkinson disease: the Norwegian ParkWest study. *JAMA Neurology*, *70*(5), 580-586.

Pedersen, K. F., Larsen, J. P., Tysnes, O. B., & Alves, G. (2017). Natural course of mild cognitive impairment in Parkinson disease: A 5-year population-based study. *Neurology*,
88(8), 767-774.

Pedraza, O., Lucas, J. A., Smith, G. E., Willis, F. B., Graff-Radford, N. R., Ferman, T. J., ..., Ivnik, R. J. (2005). Mayo's older African American normative studies: confirmatory factor analysis of a core battery. *Journal of the International Neuropsychological Society*, *11*, 184-191.

Petersen, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, *256*(3), 183-194.

Phaf, R. H., Horsman, H. H., van der Moolen, B., Roos, Y. B., & Schmand, B. (2010). A slow component of classic Stroop interference. *European Journal of Cognitive Psychology*, *22*, 306-320.

Poldrack, R. A., & Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nature Neuroscience*, *17*(11), 1510-1517.

Poline, J. B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., ... & Poldrack, R. A. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics,* 6-9.

R Code Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Rabbitt, P. (1979). How old and young subjects monitor and control responses for accuracy and speed. *British Journal of Psychology*, *70*, 305-311.

Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. *Archives of Clinical Neuropsychology*, *31*, 206-230.

Reitan, R.M. (1992). *Trail Making Test: Manual for administration and scoring.* Tucson, AZ: Reitan Neuropsychological Laboratory.

Reppermund, S., Sachdev, P. S., Crawford, J., Kochan, N. A., Slavin, M. J., Kang, K., ... & Brodaty, H. (2011). The relationship of neuropsychological function to instrumental activities of daily living in mild cognitive impairment. *International Journal of Geriatric Psychiatry*, *26*(8), 843-852. doi:10.1002/gps.2612

Revelle, W. (2008). psych: Procedures for personality and psychological research (R package version 1.7.8).

Rey, A. (1958). *L'examen clinique en psychologie.* Paris, France: Presses Universitaires de France.

Ricarte, J. J., Ros, L., Latorre, J. M., Muñoz, M. D., Aguilar, M. J., & Hernandez, J. V. (2016). Role of anxiety and brooding in specificity of autobiographical recall. *Scandinavian Journal of Psychology*, *57*(6), 495-500.

Royall, D.R., Cordes, D.A., & Polk, M. (1998). CLOX: An executive clock drawing task. *Journal of Neurology, Neurosurgery, and Psychiatry, 64*, 588-594.

Royall, D. R., Bishnoi, R. J., & Palmer, R. F. (2015). Serum IGF-BP2 strongly moderates age's effect on cognition: a MIMIC analysis. *Neurobiology of Aging*, *36*(7), 2232-2240. doi:10.1016/j.neurobiolaging.2015.04.003

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*, 87-94.

de Ruiter, M. B., Reneman, L., Boogerd, W., Veltman, D. J., van Dam, F. S., Nederveen, A. J., ... & Schagen, S. B. (2011). Cerebral hyporesponsiveness and cognitive impairment 10 years after chemotherapy for breast cancer. *Human Brain Mapping*, *32*(8), 1206-1219.

Sakia, R. (1992). The box-cox transformation technique: A review. *The Statistician*, *41*(2), 169-178

Santangelo, G., Vitale, C., Picillo, M., Moccia, M., Cuoco, S., Longo, K., ..., Amboni, M. (2015). Mild Cognitive Impairment in newly diagnosed Parkinson's disease: A longitudinal prospective study. *Parkinsonism & Related Disorders*, *21*(10), 1219-1226.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Schaefer, J., Giangrande, E., Weinberger, D. R., & Dickinson, D. (2013). The global cognitive impairment in schizophrenia: Consistent over decades and around the world. *Schizophrenia Research*, *150*(1), 42-50.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*, 23-74.

Schmand, B., de Bruin, E., de Gans, J., & van de Beek, D. (2010). Cognitive functioning and quality of life nine years after bacterial meningitis. *Journal of Infection*, *61*, 330-334.

Schmidt, C. S., Schumacher, L. V., Römer, P., Leonhart, R., Beume, L., Martin, M., ... & Kaller, C. P. (2017). Are semantic and phonological fluency based on the same or distinct sets of cognitive processes? Insights from factor analyses in healthy adults and stroke patients. *Neuropsychologia*, *99*, 148-155.

Schretlen, D. J., Peña, J., Aretouli, E., Orue, I., Cascella, N. G., Pearlson, G. D., & Ojeda, N. (2013). Confirmatory factor analysis reveals a latent cognitive structure common to bipolar disorder, schizophrenia, and normal controls. *Bipolar disorders*, *15*(4), 422-433.

Selnes, O. A., Gottesman, R. F., Grega, M. A., Baumgartner, W. A., Zeger, S. L., & McKhann, G. M. (2012). Cognitive and neurologic outcomes after coronary-artery bypass surgery. *New England Journal of Medicine*, *366*(3), 250-257.

Siedlecki, K. L., Honig, L. S., & Stern, Y. (2008). Exploring the structure of a neuropsychological battery across healthy elders and those with questionable dementia and Alzheimer's disease. *Neuropsychology*, *22*, 400.

Siedlecki, K. L., Manly, J. J., Brickman, A. M., Schupf, N., Tang, M. X., & Stern, Y. (2010). Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers? *Neuropsychology*, *24*(3), 402-411.

Small, G. W., Rabins, P. V., Barry, P. P., Buckholtz, N. S., DeKosky, S. T., Ferris, S. H., ... & McRae, T. D. (1997). Diagnosis and treatment of Alzheimer disease and related disorders: consensus statement of the American Association for Geriatric Psychiatry, the Alzheimer's Association, and the American Geriatrics Society. *JAMA*, *278*(16), 1363-1371.

Smeding, H. M., Speelman, J. D., Huizenga, H. M., Schuurman, P. R., & Schmand, B. (2011). Predictors of cognitive and psychosocial outcome after STN DBS in parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *82*, 754–760

Smeding, H. M. M., Speelman, J. D., Koning-Haanstra, M., Schuurman, P. R., Nijssen, P., Van Laar, T., & Schmand, B. (2006). Neuropsychological effects of bilateral STN stimulation in Parkinson disease: A controlled study. *Neurology*, *66*(12), 1830-1836.

Snitz, B. E., Yu, L., Crane, P. K., Chang, C. C. H., Hughes, T. F., & Ganguli, M. (2012). Subjective cognitive complaints of older adults at the population level: an item response theory analysis. *Alzheimer disease and associated disorders*, *26*(4), 344-351.

Snow, W. G. (1987). Standardization of test administration and scoring criteria: Some shortcomings of current practice with the Halstead-Reitan test battery. *The Clinical Neuropsychologist*, *1*, 250-262.

Spencer, S., & Huh, L. (2008). Outcomes of epilepsy surgery in adults and children. *The Lancet Neurology*, *7*(6), 525-537.

Steenbergen, M. R., & Jones, B. S. (2002). Modeling multilevel data structures. *American Journal of Political Science*, *46*(1), 218-237.

Sternäng, O., Lövdén, M., Kabir, Z. N., Hamadani, J. D., & Wahlin, Å. (2016). Different context but similar cognitive structures: Older adults in rural Bangladesh. *Journal of Cross-Cultural Gerontology*, *31*(2), 143-156.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. New York, NY: Oxford University Press.

Su, T., Schouten, J., Geurtsen, G. J., Wit, F. W., Stolte, I. G., Prins, M., ... & AGEhIV Cohort Study Group (2015). Multivariate normative comparison, a novel method for more reliably detecting cognitive impairment in HIV infection. *AIDS*, *29*, 547-557.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Boston: Pearson Education.

Tabert, M. H., Manly, J. J., Liu, X., Pelton, G. H., Rosenblum, S., Jacobs, M., ... & Devanand, D. P. (2006). Neuropsychological prediction of conversion to Alzheimer disease in patients with mild cognitive impairment. *Archives of General Psychiatry*, *63*(8), 916-924.

Testa, S. M., Winicki, J. M., Pearlson, G. D., Gordon, B., & Schretlen, D. J. (2009). Accounting for estimated IQ in neuropsychological test performance with regression-based techniques. *Journal of the International Neuropsychological Society*, *15*, 1012–1022.

Thibeau, S., McFall, G. P., Wiebe, S. A., Anstey, K. J., & Dixon, R. A. (2016). Genetic factors moderate everyday physical activity effects on executive functions in aging: Evidence from the Victoria Longitudinal Study. *Neuropsychology*, *30(1)*, 6-17.

Tierney, M. C., Szalai, J. P., Snow, W. G., Fisher, R. H., Nores, A., Nadon, G., ..., St. George-Hyslop, P. S. (1996). Prediction of probable Alzheimer's disease in memory-impaired patients: A prospective longitudinal study. *Neurology*, *46*, 661-665.

Tractenberg, R. E., Fillenbaum, G., Aisen, P. S., Liebke, D. E., Yumoto, F., & Kuchibhatla, M. N. (2010). What the CERAD battery can tell us about executive function as a higher-order cognitive faculty. *Current Gerontology and Geriatrics Research*, *510614*, 1-10.

Troendle, J., F. (1995). A stepwise resampling method of multiple hypothesis testing. *Journal of the American Statistical Association, 90*(429), 370-378.

Tse, C. S., Balota, D. A., Yap, M. J., Duchek, J. M., & McCabe, D. P. (2010). Effects of healthy aging and early stage dementia of the

Alzheimer's type on components of response time distributions in three attention tasks. *Neuropsychology*, *24*(3), 300-315.

Tsuji, H., Venditti, F. J., Manders, E. S., Evans, J. C., Larson, M. G., Feldman, C. L., & Levy, D. (1996). Determinants of heart rate variability. *Journal of the American College of Cardiology*, *28*(6), 1539-1546.

Tuokko, H. A., Chou, P. H. B., Bowden, S. C., Simard, M., Ska, B., & Crossley, M. (2009). Partial measurement equivalence of French and English versions of the Canadian Study of Health and Aging neuropsychological battery. *Journal of the International Neuropsychological Society*, *15*(3), 416-425.

United Nations Educational, Scientific and Cultural Organization. (2011). *International Standard Classification of Education-ISCED 2011: December 2012*. Paris, France: Author.

Valdés-Sosa, M., Bobes, M. A., Quiñones, I., Garcia, L., Valdes-Hernandez, P. A., Iturria, Y., ..., Asencio, J. (2011). Covert face recognition without the fusiform-temporal pathways. *Neuroimage*, *57*, 1162–1176.

Valenzuela, M. J., & Sachdev, P. (2007). Assessment of complex mental activity across the lifespan: development of the Lifetime of Experiences Questionnaire (LEQ). *Psychological Medicine*, *37*(7), 1015-1025.

Van der Laan, M. J., Dudoit, S., & Pollard, K. S. (2004). Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology, 3*(1), 1-33.

de Vent, N. R., Agelink van Rentergem, J. A., Schmand, B. A., Murre, J. M. J., ANDI Consortium & Huizenga, H. M. (2016a). Advanced Neuropsychological Diagnostics Infrastructure (ANDI): A normative database created from control datasets, *Frontiers in Psychology*, *7*(1601), 1-10.

de Vent, N. R., Agelink van Rentergem, J. A., Kerkmeer, M. C., Huizenga, H. M., Schmand, B. A., & Murre, J. M. J. (2016b). Universal Scale of Intelligence Estimates (USIE): Representing intelligence estimated from level of education. *Assessment*, 1-7.

Verhage, F. (1964). *Intelligentie en leeftijd onderzoek bij Nederlanders van twaalf tot zevenenzeventig jaar [Intelligence and age research with Dutch people aged twelve to seventyseven years]* (Doctoral dissertation). Assen, the Netherlands: Van Gorcum Prakke en Prakke.

Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analyses of age–cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin*, *122*(3), 231-249.

Verhoeven, K. J., Simonsen, K. L., & McIntyre, L. M. (2005). Implementing false discovery rate control: Increasing your power. *Oikos*, *108*(3), 643-647.

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., ... & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology, 24*(1), 94-97.

Voncken, L., Albers, C. J., & Timmerman, M. E. (2017). Model selection in continuous test norming with GAMLSS. *Assessment*, 1-18.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*, 228-243.

Waldinger, R. J., Cohen, S., Schulz, M. S., & Crowell, J. A. (2015). Security of attachment to spouses in late life: Concurrent and prospective links with cognitive and emotional well-being. *Clinical Psychological Science*, *3*(4), 516-529.

Watts, A. S., Loskutova, N., Burns, J. M., & Johnson, D. K. (2013). Metabolic syndrome and cognitive decline in early Alzheimer's disease and healthy older adults. *Journal of Alzheimer's Disease*, *35*(2), 253-265.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale - Revised (WAIS-R).* New York: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale (WAIS-III) (3rd ed.).* New York: Psychological Corporation

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and*

methods for p-value adjustment (Vol. 279). John Wiley & Sons.

Wettstein, M., Kuźma, E., Wahl, H. W., & Heyl, V. (2016). Cross-sectional and longitudinal relationship between neuroticism and cognitive ability in advanced old age: The moderating role of severe sensory impairment. *Aging & Mental Health*, *20*(9), 918-929.

Whittle, C., Corrada, M. M., Dick, M., Ziegler, R., Kahle-Wrobleski, K., Paganini-Hill, A., Kawas, C. (2007). Neuropsychological data in nondemented oldest old: the 90+ study. *Journal of Clinical and Experimental Neuropsychology*, *29*, 290–299.

Williams, P. G., Suchy, Y., & Kraybill, M. L. (2010). Five-factor model personality traits and executive functioning among older adults. *Journal of Research in Personality*, *44*(4), 485-491.

Williams-Gray, C. H., Foltynie, T., Brayne, C. E. G., Robbins, T. W., & Barker, R. A. (2007). Evolution of cognitive dysfunction in an incident Parkinson's disease cohort. *Brain*, *130*(7), 1787-1798.

Wilson, B., Cockburn J., & Baddeley, A. (1983). *Rivermead Behavioural Memory Test.* Reading, UK: Thames Valley Test Company.

12

FUNDING, RESULTING ARTICLES AND SHORT CV

The ANDI project was funded by MaGW grant 480-12-015.

Chapter 2 published as:

de Vent, N. R.*, Agelink van Rentergem, J. A.*, Schmand, B. A., Murre, J. M. J., ANDI Consortium & Huizenga, H. M. (2016). Advanced Neuropsychological Diagnostics Infrastructure (ANDI): A normative database created from control datasets. *Frontiers in Psychology*, *7*(1601), 1-10.

NRdV developed the method, assembled the database, wrote the article. JAvR developed the method, analyzed the data, wrote the article. BS developed the concept, supervised data collection, revised the article. JM developed the concept, revised the article. ANDI consortium provided data. HH developed the concept, supervised the analysis, revised the article.

Chapter 3 published as:

Agelink van Rentergem, J. A., Murre, J. M. J., & Huizenga, H. M. (2017). Multivariate normative comparisons using an aggregated database. *PLoS ONE*, *12*, 1-18.

JAvR developed the method, analyzed the data, wrote the article. JM developed the concept, revised the article. HH developed the concept, supervised the analysis, revised the article.

Chapter 4 published as:

Agelink van Rentergem, J. A., de Vent, N. R., Schmand, B. A., Murre, J. M. J., & Huizenga, H. M. (2017). Multivariate normative comparisons for neuropsychological assessment by a multilevel factor structure or multiple imputation approach, *Psychological Assessment*. Advance online publication. doi:10.1037/pas0000489

JAvR developed the method, analyzed the data, wrote the article. NRdV assembled the database, revised the article. BS developed the concept, supervised data collection, revised the article. JM developed the concept, revised the article. HH developed the concept, supervised the analysis, revised the article.

Chapter 5 submitted as:

Agelink van Rentergem, J. A., de Vent, N. R., Schmand, B. A., Murre, J. M. J., Staaks, J. P. C., ANDI Consortium, & Huizenga, H. M. (2017). *Cognitive domains in neuropsychology: Support for the Cattell-Horn-Carroll model in two research syntheses*. Manuscript submitted for publication.

JAvR developed the concept, developed the method, analyzed the data, wrote the article. NRdV assembled the database, revised the article. BS supervised data collection, revised the article. JM revised the article. JS supervised literature search, revised the article. ANDI consortium provided data. HH developed the concept, supervised the analysis, revised the article.

Chapter 6 submitted as:

Agelink van Rentergem, J. A.*, de Vent, N. R.*, Huizenga, H. M., Murre, J. M. J., ANDI Consortium, & Schmand, B. A. (2017). *Predicting Parkinson's disease dementia using modern neuropsychological techniques*. Manuscript submitted for publication.

JAvR developed the method, analyzed the data, wrote the article. NRdV developed the method, assembled the database, wrote the article. HH supervised the analysis, revised the article. JM revised the article. ANDI consortium provided data. BS developed the concept, provided data, supervised data collection, revised the article.

Chapter 7 published as:

Zadelaar, J. N.*, Agelink van Rentergem, J. A.*, & Huizenga, H. M. (2017). Univariate comparisons given aggregated normative data. *The Clinical Neuropsychologist*, *31*(6-7), 1155-1172.

JZ developed the method, analyzed the data, wrote the article. JAvR developed the concept, developed the method, analyzed the data, revised the article. HH developed the concept, supervised the analysis, revised the article.

Other publications, not included in this thesis:

Dekkers, T. J., Popma, A., Agelink van Rentergem, J. A., Bexkens, A., & Huizenga, H. M. (2016). Risky decision making in Attention-Deficit/Hyperactivity Disorder: A meta-regression analysis. *Clinical Psychology Review*, *45*, 1-16.

Menning, S., de Ruiter, M. B., Kieffer, J. M., Agelink van Rentergem, J., Veltman, D. J., Fruijtier, A., ... Schagen, S. B. (2016). Cognitive impairment in a subset of breast cancer patients after systemic therapy—Results from a longitudinal study. *Journal of Pain and Symptom Management*, *52*, 560-569.

Huizenga, H. M., Agelink van Rentergem, J. A., Grasman, R. P. P. P., Muslimovic, D., & Schmand, B. (2016). Normative comparisons for large neuropsychological test batteries: User-friendly and sensitive solutions to minimize familywise false positives. *Journal of Clinical and Experimental Neuropsychology*, *38*, 611-29.

de Vent, N. R., Agelink van Rentergem J. A., Kerkmeer M. C., Huizenga H. M., Schmand B. A., & Murre J. M. J. (2016). Universal Scale of Intelligence Estimates (USIE): Representing intelligence estimated from level of education. *Assessment*, 1-7.

Wagemaker, E.*, Dekkers, T. J.*, Agelink van Rentergem, J. A., Volkers, K. M., & Huizenga, H. M. (2017). Advances in mental health care: Five N= 1 Studies on the effects of the robot seal Paro in adults with severe intellectual disabilities. *Journal of Mental Health Research in Intellectual Disabilities*, 1-12.

Dekkers, T. J., Agelink van Rentergem, J. A., Koole, A., van den Wildenberg, W. P. M., Popma, A., Bexkens, A., ... Huizenga, H. M. (2017). Time-on-task effects in children with and without ADHD: depletion of executive resources or depletion of motivation? *European Child & Adolescent Psychiatry*, 1-11.

Dekkers, T. J., Agelink van Rentergem, J. A., Huizenga, H. M., Raber, H., Shoham, R., Popma, A., & Pollak, Y. (2017). *Decision-making deficits in Attention-Deficit/Hyperactivity Disorder (ADHD) are not related to risk seeking but to suboptimal decision making: meta-analytical and novel experimental evidence*. Manuscript submitted for publication.

Dekkers, T. J.*, Agelink van Rentergem, J. A., Meijer, B., Popma, A., & Huizenga, H. M. (2017). *A meta-analytical evaluation of the dual-hormone hypothesis: Does cortisol moderate the relationship between testosterone and status-relevant behavior?* Manuscript submitted for publication.

Other publications

Dovis, S., Agelink van Rentergem, J., & Huizenga, H. M. (2015). Does Cogmed Working Memory Training Really Improve Inattention in Daily Life? A Reanalysis [Letter to the editor]. *PLoS One*, *10*(3).

Dovis, S., Agelink van Rentergem, J., & Huizenga, H. M. (2015). Response to the Correction by Spencer-Smith and Klingberg: Unaddressed Concerns [Letter to the editor]. *PLoS One*, *10*(3).

Dovis, S., Agelink van Rentergem, J., & Huizenga, H. M. (2015). Does brain training really help ADHD? *New Scientist*.

Dovis, S., Agelink van Rentergem, J., & Huizenga, H. M. (2016). Concerns about the Corrected Review and Meta-Analysis of Cogmed Working Memory Training Effects on Inattention in Daily Life. [Letter to the editor]. *PLoS One*, *10*(3).

Dekkers, T. J., Agelink van Rentergem, J. A., Popma, A., Bexkens, A., & Huizenga, H. M. (2016). Risicovol beslisgedrag bij ADHD: een meta-regressie analyse. *Neuropraxis*.

Schmand, B. A., Agelink van Rentergem, J. A., de Vent, N. R., Murre, J. M. J., & Huizenga, H. M. (2017). Advanced Neuropsychological Diagnostics Infrastructure (ANDI). Voor een scherpere neuropsychologische diagnostiek. *Tijdschrift voor Neuropsychologie*.

*authors contributed equally

Joost Agelink van Rentergem was born as Joost Zandvliet, in Amsterdam, on 23rd of December 1987. After the 10th Montessori school “De Meidoorn”, and the Montessori Lyceum Amsterdam, he studied Psychology at the University of Amsterdam, obtaining both his Bachelor´s and Research Master´s degrees cum laude. The research for this PhD project was conducted between September 2013 and November 2017, and was supervised by prof.dr. Hilde Huizenga, prof.dr. Ben Schmand and prof.dr. Jaap Murre. This research was conducted in close collaboration with Nathalie Ramona de Vent.

13

NEDERLANDSE SAMENVATTING

Het doel van dit proefschrift was om de betrouwbaarheid van neuropsychologisch onderzoek te vergroten, door de normatieve vergelijkingsprocedure te verbeteren. Het eerste doel was om multivariate normatieve vergelijkingen mogelijk te maken, die het hele profiel van een patiënt toetsen. Het tweede doel was om normatieve vergelijkingen mogelijk te maken die gecorrigeerd zijn voor leeftijd, sekse en opleiding. Er waren twee voorwaarden voor het behalen van deze doelen. In de eerste plaats moest er een normatieve database worden opgesteld met veel, demografisch diverse, gezonde deelnemers. Daarnaast moesten er statistische methodes ontwikkeld worden voor het maken van multivariate normatieve vergelijkingen met deze normatieve database. Deze statistische methoden vormden het onderwerp van dit proefschrift. In hoofdstuk twee hebben we beschreven hoe een geaggregeerde normatieve database kan worden gebouwd door data van gezonde personen uit meerdere studies te combineren. Deze mensen kunnen hebben deelgenomen als proefpersonen in een controlegroep in een klinische studie, of kunnen hebben meegedaan aan een groot bevolkingsonderzoek. Door veel van zulke groepen te combineren, kunnen veel data van verschillende neuropsychologische taken worden verzameld. De procedures werden gestandaardiseerd voor de verschillende tests. Dit hield twee procedures in voor het opschonen van de data. Ten eerste werden waarden verwijderd die buiten een vooraf gedefinieerd bereik van toelaatbare scores lagen dat vooraf was vastgesteld op basis van klinische expertise. Ten tweede werden waarden verwijderd die zeer onwaarschijnlijk waren gegeven de leeftijd, sekse en opleiding van de deelnemers. Om te bepalen welke demografische variabelen zouden worden gebruikt in de demografische correcties, werd gebruik gemaakt van het Akaike Informatie Criterium. Om gebruik te kunnen maken van parametrische statistiek zoals parametrische normatieve vergelijkingen, zouden scores idealiter normaal verdeeld zijn, of getransformeerd moeten worden zodat zij normaal verdeeld zijn. Voor de selectie van de macht tot waar de data moesten worden verheven opdat deze normaal verdeeld waren, hebben wij gebruikgemaakt van de Box-Cox procedure (Box & Cox, 1964). Tot slot wordt de inhoud van de ANDI-database ook in dit hoofdstuk beschreven.

In hoofdstuk drie hebben we beschreven hoe met een geaggregeerde database multivariate normatieve vergelijkingen kunnen worden gemaakt. Hiervoor is een model nodig dat bestaat uit drie gedeelten. Ten eerste was er om demografische correcties uit te kunnen voeren voor leeftijd, geslacht en opleidingsniveau een regressiemodel nodig, om regressiecoëfficiënten te kunnen schatten voor deze drie demografische variabelen. Ten tweede zouden er verschillen kunnen bestaan tussen studies in de scores die gezonde proefpersonen behalen, bijvoorbeeld door verschillen tussen studies in hoe de steekproef geselecteerd was, of hoe de testen werden afgenomen. Daarom was een multilevel model nodig, om de verschillen tussen studies te modelleren. Ten derde houden multivariate normatieve vergelijkingen rekening met de relaties tussen scores op verschillende tests. Om de covariantie tussen scores te schatten was een multivariaat model nodig. Om deze onderdelen samen te voegen, werd een multivariaat multilevel regressiemodel geformuleerd. Dit multivariate multilevel regressiemodel heeft als bijkomend voordeel dat het ook kan worden toegepast wanneer er ontbrekende waardes zijn binnen de testvariabelen. Vanwege de geaggregeerde structuur van de database is een grote hoeveelheid ontbrekende waardes te verwachten, omdat tests die niet zijn afgenomen in een bepaalde studie alleen maar ontbrekende waardes hebben voor de deelnemers in deze studie. Met behulp van het multivariate multilevel regressiemodel kunnen alle componenten die nodig zijn voor multivariate normatieve vergelijkingen worden geschat: de demografisch gecorrigeerde gemiddelden, de varianties en de covarianties. In een simulatiestudie werden de prestaties van de multivariate normatieve vergelijkingenprocedure geëvalueerd, met verschillende hoeveelheden ontbrekende waardes en tussenstudievariantie. Gevonden werd dat hoewel het model kan worden toegepast met ontbrekende waardes, dit niet mogelijk is als er ontbrekende overlap is tussen tests. Dit probleem wordt behandeld in hoofdstuk vier.

In hoofdstuk vier hebben we beschreven hoe het model uit hoofdstuk drie kan worden uitgebreid om ontbrekende overlap tussen tests op te vangen. Er ontbreekt overlap tussen twee tests als de combinatie van deze twee tests nooit is afgenomen in een van de studies die zijn opgenomen in de database. Dit maakt het onmogelijk om de covariantie tussen deze twee tests direct te schatten. In dit hoofdstuk worden twee methoden behandeld die dit probleem zouden kunnen oplossen. De eerste is multipele imputatie, waarmee waardes worden ingevuld voor elke ontbrekende waarde. Met deze ingevulde waardes kan de covariantie op een eenvoudige manier worden geschat. De tweede is een factormodelbenadering, waarbij een model voor de covariantiestructuur wordt geschat. Dit model gaat ervan uit dat de covariantie tussen tests kan worden beschreven door middel van de afhankelijkheid van deze tests op eenzelfde latente variabele. In een simulatieonderzoek werden de twee methoden vergeleken. De multipele imputatiebenadering houdt het aantal fout-positieven onder controle, maar vanwege onderschatting van de covariantie tussen tests is zij minder gevoelig voor het detecteren van daadwerkelijke afwijkingen dan de factormodelbenadering. Een vereiste voor de factormodelbenadering is de geschiktheid van het factormodel voor de data. Als het factormodel niet past neemt het aantal fout-positieven toe. Daarom moet een factormodel voor neuropsychologische tests worden vastgesteld voordat deze benadering kan worden toegepast. Dit probleem wordt in hoofdstuk vijf behandeld.

In hoofdstuk vijf wordt, in twee studies, vergeleken hoe verschillende factormodellen voor neuropsychologische tests passen. Voor de eerste studie, een meta-analyse, zijn de correlatiematrices van neuropsychologische tests opgevraagd van gepubliceerde studies. De correlatie van de testscores met demografische variabelen werd uit de correlatie tussen tests verwijderd. Vervolgens zijn de correlatiematrices samengevoegd in een enkele correlatiematrix, waarop factormodellen kunnen worden gepast. In de tweede studie zijn factormodellen gepast op demografisch gecorrigeerde data uit de ANDI-database. In beide studies wordt door middel van modelvergelijkingen aangetoond dat het Cattell-Horn-Carroll-model zoals aangepast door Jewsbury et al. (2016) het beste past. Dit model is oorspronkelijk ontwikkeld in intelligentieonderzoek. Het verdeelt het cognitief functioneren zoals gemeten door neuropsychologische tests in domeinen van "Verworven kennis of gekristalliseerde vaardigheid", "Verwerkingssnelheid", "Encoderen en ophalen bij langetermijngeheugen", "Werkgeheugen", en "Woordfluency". Dit is in tegenstelling tot andere modellen die cognitief functioneren verdelen in domeinen van "Aandacht", "Executief functionerenën "Geheugen". Omdat het model van Cattell-Horn-Carroll goed lijkt te passen op data van gezonde mensen, kan dit model in ANDI worden gebruikt om de methoden uit hoofdstuk vier toe te passen.

In hoofdstuk zes worden de in dit proefschrift ontwikkelde methoden empirisch getoetst. De ANDI-database en multivariate normatieve vergelijkingen zijn gebruikt in een heranalyse van longitudinale gegevens van een onderzoek naar de ziekte van Parkinson en dementie (Broeders et al., 2013). Deze data zijn eerder geanalyseerd met de conventionele (univariate) criteria voor milde cognitieve stoornissen bij de ziekte van Parkinson (PD-MCI, Parkinson's Disease-Mild Cognitive Impairment; Litvan et al., 2012). Het doel van de studie van Broeders et al. (2013) was om te onderzoeken of degenen die bij de eerste meting aan de PD-MCI-criteria voldeden, op een later meetmoment dementie zouden hebben ontwikkeld. In dit hoofdstuk werden de resultaten van deze studie vergeleken met resultaten verkregen met de ANDI-database. Ten eerste leverde de toepassing van de univariate PD-MCI-criteria met de ANDI-database voorzichtiger resultaten op dan het eerdere onderzoek: bij minder patiënten werd een milde cognitieve stoornis vastgesteld. Dit was het geval bij zowel patiënten die later dementie ontwikkelden als bij patiënten die geen dementie ontwikkelden. Ten tweede blijken multivariate normatieve vergelijkingen met de ANDI-database betere voorspellingen te bieden dan de conventionele PD-MCI-criteria: ze zijn zowel meer sensitief als meer specifiek in het voorspellen van wie dementie ontwikkelt. Dit wijst erop dat de hier beschreven methoden nuttig zijn bij het verbeteren van neuropsychologisch onderzoek.

In hoofdstuk zeven keren we terug naar het probleem van het gebruik van univariate normatieve vergelijkingen in de klinische neuropsychologie. Als er geen correctie wordt gebruikt en er veel univariate normatieve vergelijkingen worden uitgevoerd voor veel verschillende testvariabelen, neemt het aantal keren toe dat cognitie bij gezonde mensen ook als afwijkend wordt beschouwd. Dit houdt in dat er een verhoogd fout-positievenpercentage bestaat voor een groep statistische toetsen. Dit kan hebben bijgedragen aan de lagere specificiteit voor de PD-MCI-criteria in hoofdstuk zes. Om voor dit toegenomen fout-positievenpercentage te corrigeren, zijn correctiemethoden ontwikkeld. Een correctiemethode die in de wetenschap veel wordt gebruikt, maar niet zozeer in de klinische praktijk, is de Bonferroni-correctie. Deze correctie verlaagt het aantal fout-positieven, maar kan de kans schaden dat daadwerkelijke afwijkingen worden opgespoord. In dit hoofdstuk worden verfijndere correctiemethoden besproken en vergeleken in een simulatieonderzoek, specifiek voor de situatie waarin patiënten worden vergeleken met een geaggregeerde database. Een nieuwe stapsgewijze methode presteerde in veel gevallen beter dan de Bonferroni-correctie bij het detecteren van stoornissen, maar vertoonde wel een toename van fout-positieven als veel data ontbraken. Daarom is het te vroeg om één van de methoden als de beste aan te wijzen.

Dit proefschrift ging gepaard met de bouw van de ANDI-database en -website. Voor dit project hebben onderzoeksgroepen uit Nederland en België ruimhartig data geschonken van 27.000 deelnemers. In het ANDI-project zijn de methodes gebruikt die zijn beschreven in hoofdstuk twee en drie. De website zal nog worden uitgebreid met de methode die in hoofdstuk vier is beschreven, gebruikmakend van het model dat in hoofdstuk vijf is beschreven.

14

ACKNOWLEDGEMENTS - DANKWOORD

Hilde, eeuwig dank voor de afgelopen vijf jaar. Onder jouw hoede ben ik gegroeid, omdat je me altijd een voedingsbodem en bescherming bood. En deze plantenmetafoor is niet geheel toevallig. Ben, het was een eer om bij jou te mogen promoveren, wat helemaal niet klopt met hoe zorgzaam en gezellig het voelde. Jaap, met plezier heb ik met je samengewerkt, en ik heb zo het gevoel dat dat nog wel even blijft.

Roy, Marieke, Frans, Conor, ik ben zo blij dat jullie in mijn commissie wilden. Heel veel dank.

Sanne en Hilde, dank dat jullie in mijn commissie plaats hebben willen nemen, en heel veel dank voor jullie vertrouwen. Ik werk met bijzonder veel zin en plezier in het werk met jullie samen.

Nathalie, work spouse, het klikte meteen, en we hebben vervolgens vier jaar lang gekibbeld. Het is wennen om niet meer dagelijks aan je keiharde humor blootgesteld te worden ;-). Ik ben trots op ons.

Django, paranimf, dank voor je tientallen jaren aan vriendschap. Ik zou een ander, ongelukkiger, slechter, mens geweest zijn zonder jou.

Tycho, seriële co-auteur, dank dat je me in zoveel dingen hebt betrokken en gesteund. Daar zijn de afgelopen vier jaar zoveel leuker en rijker door geworden.

Jacqueline, mijn oud-student en collega, dank dat je het allemaal zo goed hebt gedaan. Een these, een AiO-plek, een keertje submitten, en nu staat je eerste artikel in dit proefschrift, veel dank!

Laura, dank voor de regelmatige avonden samen werken, je goede zorgen, al de thee en dat ik je altijd mocht storen als ik weer eens een verhaal wilde vertellen, heel lief.

Rooske, ik geloof dat jij en ik het over alles eens zijn, en we hebben inmiddels wat onderwerpen gehad. Heel veel dank voor je vriendschap.

Helen, zo lang hebben we niet samengewerkt, maar toch was je meteen een steun en toeverlaat. Dank je wel.

Riëtte en Eline, dank voor de gezellige samenwerking op de twee papers met Hilde.

Daan, Maaike, jullie zijn als kamergenoten zo stevig gebleken, dat ik jullie nooit ergens mee heb mogen helpen. Het gaat jullie goed, en dat zal zo blijven.

Janneke, dank voor de zoekskills die je me hebt geleerd. Ik heb er veel baat bij gehad, en heb dat nog steeds.

ANDI consortium, dank jullie wel! ANDI stond of viel natuurlijk met jullie data, dus veel dank voor jullie bijdrage en jullie hulp.

eScience center, thank you for creating the ANDI website! It has been good working together with you, and I learnt a lot. This PhD project was better for it. Mateusz, Janneke, Anand, Lars, thank you.

Bas, onze samenwerking had bepaald een wervelwindkwaliteit, en dat lag meer aan jou dan aan mij. Dank dat je mij in dat projectje hebt willen betrekken.

Mijke, Raoul, Dora, Dylan, Robert, veel dank voor jullie meedenken met mijn methodologische issues, het heeft echt geholpen.

Dear students, beste studenten, thank you for listening and watching me doodle. I do hope some of it made sense. Robin, Carlijn, Odette en Liza, veel dank dat jullie mijn eerste en meteen favoriete bachelorgroep hebben willen zijn.

Brenda, dank voor de mooie open manier waarmee je mij altijd tegemoet bent getreden. Helle, je kreeg als mede-organisator een totaal onorganisatorisch type, maar dankzij je geduld en liefheid bleek het samen organiseren van Rita Vuyk een leuke klus, waarvoor veel dank. Annematt, dank voor je enthousiasme en je stimulerende gesprekken, waarvan ik verwacht dat er nog wel zullen volgen. Patrick, dank voor je verhalen, en natuurlijk dank voor het verbeteren van mijn Engels, waarbij je mijn trots ook hebt weten te sparen. Kiki, it's been a while, but I do remember fondly our conversations back at the Diamantbeurs. Bram, dank voor alle praatjes die we hebben gemaakt als we elkaar "toevallig" tegenkwamen. Gorka, live long and prosper. Ellen, dank voor alle gezelligheid en hulp, en ons tweemansmuseumclubje. Helma, Eveline, Hubert, dank voor de hulp bij onmogelijke verzoeken. Annemie, dank voor de gezelligheid, en het onverwachte inspirerende boek! Tjitske, dank dat ik (en de rest van de afdeling) op je heb mogen bouwen. Er ontbreken nog veel mensen, a lot of people are still missing from this list. Thank you all for the time I have had the past four years.

Conor, Lourens, Harrie, Jelte, Denny, Philippe, Renée, Ineke, jullie hebben mij op de cruciale momenten op de basisschool, middelbare school en universiteit, aan mijn haren uit het moeras getrokken. Ik kan dus ook met alle zekerheid zeggen dat dit proefschrift er zonder jullie niet geweest was, want zonder jullie hulp was ik er nooit aan begonnen. Veel dank! Eveline, zonder jouw ontzettend lieve steun en tips was dit allemaal ook nooit wat geworden.

Tsjangis, Luitzen, Ruben, Santi, Joeri, dank voor jullie vriendschap. En de rest van de familie en vriendenclub natuurlijk ook.

LLLT.

ইন্দুশ্রী

Met vriendelijke groet,
Joost