Advanced neuropsychological diagnostics infrastructure

Improving neuropsychology

de Vent, N.R.

Publication date
2021

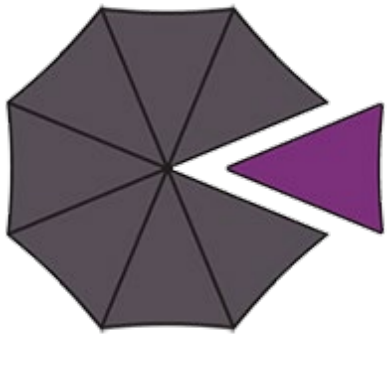
Document Version
Final published version

License
Other

Advanced Neuropsychological Diagnostics Infrastructure

Improving Neuropsychology

Nathalie Ramona de Vent



ADVANCED NEUROPSYCHOLOGICAL DIAGNOSTICS INFRASTRUCTURE

IMPROVING 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

op donderdag 21 januari 2021, te 13:00 uur

door Nathalie Ramona de Vent

geboren te Hoorn

Promotiecommissie:

Promotores:	prof.dr. B.A. Schmand, Universiteit van Amsterdam
	prof.dr. J.M.J. Murre, Universiteit van Amsterdam
Copromotor:	prof.dr. H.M.H. Huizenga, Universiteit van Amsterdam
Overige leden:	prof.dr. E.H.F. de Haan, Universiteit van Amsterdam
	prof.dr. S.B. Schagen, Universiteit van Amsterdam
	prof.dr. J.M. Spikman, Rijksuniversiteit Groningen
	prof.dr. R.P.C. Kessels, Radboud Universiteit Nijmegen
	dr. E van den Berg, Erasmus UMC
	dr. S.P. van der Werf, Universiteit van Amsterdam

Contents

Chapter 1

General Introduction

An important part of neuropsychological assessment is the process of assessing a patient's cognition via batteries of cognitive tests such as memory, executive functions and language tests. The goal of this assessment is to create a comprehensive overview of a patient's cognitive capacities and to detect possible abnormalities which can be indicative of a cognitive disorder. Neuropsychological assessment is not exclusively used in the clinical evaluation of a single patient, but also on a larger scale in research settings. In intervention studies it can be necessary to evaluate whether specific cognitive functions, such as memory after a memory training, have either improved or not (Rapp, Brenes, & Marsh, 2002). Similarly, neuropsychological assessment may be used to evaluate whether there are possible negative cognitive side effects from treatments such as radiation treatment or chemo-therapy in the case of cancer (Duffer, 2004; Stouten-Kemperman et al., 2015). And lastly, cognition is evaluated when investigating the effects of non-brain diseases such as type II diabetes (van den Berg et al., 2008) or HIV (Janssen, Bosch, Koopmans, & Kessels, 2015; Su et al.,2015).

In both the clinical evaluation as well as in the research setting it is of high importance that neuropsychological evaluation is reliable, meaning that it enables a correct discrimination between normal and abnormal cognition. In the clinic, this is important because the choice of the intervention depends on it. A patient with abnormal cognition who is later diagnosed with a specific brain disease can start their appropriate treatment, and patients who appear cognitively healthy might be referred to further investigate the origin of their symptoms. This differentiation is also highly relevant for a research setting as correctly identifying the patient population strengthens the results (e.g. increases the power of the study). Similarly, when a specific cognitive function (such as memory function in a group of MCI patients) is evaluated

correctly, it strengthens the findings of an intervention. Unfortunately, the process of neuropsychological assessment is not without its difficulties. These difficulties will be described in detail in paragraph 1.2.

The goal of this thesis is to help mend problems typically found in neuropsychological assessment, either in clinical evaluation or in the research setting. By improving neuropsychological assessment, a more reliable distinction of normal versus abnormal cognition can be made, which can lead to better treatment outcomes in the clinical setting and to a more representative patient sample in research. To this end, the Advanced Neuropsychological Diagnostics Infrastructure (ANDI) project was started. This thesis will describe the construction of ANDI as well as proofs of principle of its use in both clinical and research contexts. The current chapter will introduce the difficulties found in neuropsychological assessment in clinical evaluations as well as in research settings, and will give an overview of the thesis contents.

1.1 Normative comparisons

An important aspect of the neuropsychological assessment is deciding whether scores a patient has obtained are considered normal or abnormal. To do so, the scores of the patient are compared to a set of reference scores of healthy participants (norm group or normative sample). Ideally, the norm group is similar to the patient in terms of age, sex, and level of education, as these factors are known to influence performance on neuropsychological tests (Lezak et al., 2012). For example, what is deemed a normal score on a speeded test for a 70-year-old might be considered an abnormal score for a 30-year-old, as processing speed diminishes with age. Similarly, a patient with a high level of education is expected to perform better on a memory test than a patient with a lower level of education. Considering a patients demographic background is therefore necessary to make correct predictions on what scores are considered normal for any patient.

1.2 Difficulties in neuropsychology

In traditional neuropsychological assessment the reference or norm scores are found in tables that accompany neuropsychological test materials. Each neuropsychological test comes with its own set of norm scores, which are of mixed quality. First, some norm data were collected decades ago; possibly, these data are no longer an accurate representation of the patients seen in clinics today (Strauss et al., 2006). Second, many norm tables lack information about the oldest participants (80+), even though they make up a large portion of the patients seen in clinics (Whittle et al., 2007). Third, in some cases local norms are not available at all, and clinicians use normative data from other countries or norm data they have gathered themselves (Crawford and Garthswaite, 2002). Fourth, when using published norm tables, scores can often only be corrected for age and in some cases for sex, but not for level of education, even though this is known to influence performance on neuropsychological tests (Lezak et al., 2012). Fifth, when a patient becomes one year older, and shifts from one age-group to the next in the norm table, the interpretation of that test score can change. A score that was considered abnormal at age 59 might be considered normal now the participant is 60 years old (Zachery, & Gorsuch, 1985). What these problems have in common is that they decrease the ability to get an accurate representation of a patient's abilities, and thus create a problem when having to decide whether a score is considered normal or indicative of impairment. Sixth, in traditional neuropsychological assessment it is common to evaluate a patient's test results in a one-by-one (univariate) fashion. That is, for each score obtained from a patient, a separate comparison with a norm group is made. When several tests are used (which is often the case in neuropsychological assessment), the likelihood of obtaining at least one abnormal score by chance alone increases as the number of tests administered increases (Binder, Iverson, & Brooks, 2009). Last, because the norm data for most tests have been collected separately (most tests have not been co-normed), the correlations between neuropsychological tests are often unknown. However, clinicians do not interpret the test results as stand-alone tests but as part of the broader context of the neuropsychological evaluation. A patient's score on a delayed recall tests always gets interpreted within the scope of what they scored on the immediate recall of the test. Until now, this is only done intuitively.

1.3 Advanced Neuropsychological Diagnostics Infrastructure (ANDI)

To mend these problems in neuropsychological assessment the Advanced Neuropsychological Diagnostics Infrastructure (ANDI) was created. ANDI is an online portal which helps clinicians and researchers decide whether scores obtained from their patients are deemed normal or abnormal. At the core of ANDI lies a large, diverse database of reference data which were donated by neuropsychology researchers from The Netherlands and Flemish Belgium. ANDI circumvents problems typically found in traditional norms (described above). First, the norm data for ANDI were gathered with the help of the ANDI-consortium, and the majority of these norm data have been gathered in research projects over the past years, making them thus more representative for patients seen in clinics today. Second, the ANDI normative database contains data from men and women of all ages and all levels of education. This means that also the oldest age groups (80+) are well represented in ANDI. Third, because the data originate from a local consortium, clinicians and researchers can use normative data gathered in their own language area (i.e. Dutch), making the normative data more representative of their patient population. Fourth, because the ANDI database consists of a diverse set of healthy participants of whom information on age, sex, and level of education is available, patient scores can simultaneously be corrected for these demographic factors. This means that clinicians can better predict a patient's expected score, and consequently, any abnormality is more reliably detected. Fifth, ANDI uses regression-based norms, i.e. norms that are continuous and not stratified for age, sex and/or level of education. This circumvents the problem of a change in interpretation of scores when a patient becomes one year older and moves from one age-group to the next. These improvements together make for a more representative normative sample for each patient. Sixth, ANDI contains data from a large number of neuropsychological tests, which have been administered together (co-normed). This facilitates normative comparison methods that were not feasible before, particularly those who formally evaluate correlations between tests. Multivariate normative comparison (MNC) (Huizenga, et al., 2007) is a profile analysis which compares all test results acquired from a patient in one single analysis. The multivariate normative comparisons look at peaks and troughs in a patient's test profile and compares these to the profiles of the normative sample. This is more intuitive for clinicians, because they usually look at the neuropsychological

test results as a score profile and they do not interpret test results as stand-alone tests. Multivariate normative comparisons can detect abnormal combinations of test scores in a score profile, which can be overlooked in the traditional (univariate) normative comparisons (Crawford & Garthwaite, 2002; Huizenga et al., 2007; Su et al., 2015). Also, because the multivariate normative comparison entails only one statistical test, the problem of too high chances of obtaining at least one abnormal result is circumvented.

1.4 ANDI website

The ANDI-database is accompanied by an online portal which clinicians and researchers can use to upload the patient data they wish to analyze. After uploading, a detailed report of both the univariate and multivariate normative comparisons for each individual patient is returned to the user. The programming and scripts behind ANDI are open source and freely downloadable (www.andi.nl). By doing so, we hope to encourage colleagues in other countries to also start an ANDI-initiative as they would only have to gather their own reference data. Thus far we have had researchers from Indonesia, The USA, Czech Republic, and Australia who are very interested in building their own version of ANDI.

1.5 Thesis overview

In this thesis, the construction, application and future directions of ANDI will be discussed. Chapter two describes the construction of the ANDI database. It also describes the steps required to combine all donated datasets into a single, coherent database. These steps include matching variables, an extensive outlier removal procedure, determining the influence of demographic variables and, if needed, correcting for these, and lastly, finding appropriate transformations to normality in order to facilitate normative comparisons. The chapter also offers a short description of the neuropsychological tests found in ANDI.

In chapter three, an application of the ANDI concept in a sample of newly diagnosed patients with Parkinson's disease is given. Parkinson's patients are at higher risk of developing dementia and the goal of this chapter is to see if ANDI and the multivariate normative comparisons can be used to predict which patients would progress to dementia at three and five-year follow up. The results are compared to standard

Parkinson's disease Mild Cognitive Impairment criteria, which are traditionally used to diagnose cognitive impairment in Parkinson's disease.

In chapter four, ANDI is used to define 'abnormal cognition' in order to optimally predict progression to dementia in a memory clinic sample (from the Amsterdam Dementia Cohort). The literature does not agree on what constitutes 'abnormal cognition'. In this chapter, both the number of deviating scores (i.e. one, two or three deviating scores), as well as the magnitude of the deviations (i.e. -1 sd, -1.5 sd or -2 sd below the mean) are calibrated in order to find the combination which best predicts progression to dementia in this sample. Secondly, the predictive power of a profile analysis of the cognitive test results (MNC) for progression to dementia is compared to standard clinical practice.

Chapter five takes a side step. In neuropsychological assessment it is often necessary to acquire an estimate of a patient's premorbid level of cognitive functioning. Usually this is estimated by using a patient's level of education as a proxy. However, because level of education can be expressed on a large number of scales, it is difficult to compare results between scales. In this chapter the Universal Scale of Intelligence Estimates (USIE) is introduced. USIE is a new scale that has the capacity to replace existing scales and to be used to translate one education scale to another. USIE uses standard IQ scores and has estimation intervals to express the certainty of the IQ estimation.

The last chapter summarizes the results from all previous chapters and will discuss the limitations, possible solutions and directions for future research.

Chapter 2

ANDI: A Normative Database Created From Control Datasets

This chapter has been published as:

N.R. de Vent*, J.A. Agelink van Rentergem*, B.A. Schmand, J.M.J. Murre, ANDI Consortium, & H.M. Huizenga, (2016). Advanced Neuropsychological Diagnostics Infrastructure (ANDI): A Normative Database Created From Control Datasets. *Frontiers in Psychology 7*(1601), 1-10

*Shared first authorship

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.

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 data of neuropsychological tests might have become outdated and no longer represent the patients we see today (Strauss, Sherman, & Spreen, 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 and 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, Howieson, Bigler, & Tranel, 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 and Garthwaite, 2002; Huizenga, Smeding, Grasman, & Schmand, 2007; Castelli et al., 2010; Schmand, de Bruin, de Gans, & van de Beek, 2010; Smeding, Speelman, Huizenga, Schuurman, & Schmand, 2011; Valdés-Sosa et al., 2011; González-Redondo et al., 2012; Broeders et al., 2013; Cohen et al., 2015; 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). 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, 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 tests that are also included in the ANDI database. The AVLT measures memory and learning (Strauss et al., 2006; Lezak et al., 2012). 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.

2.1 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.

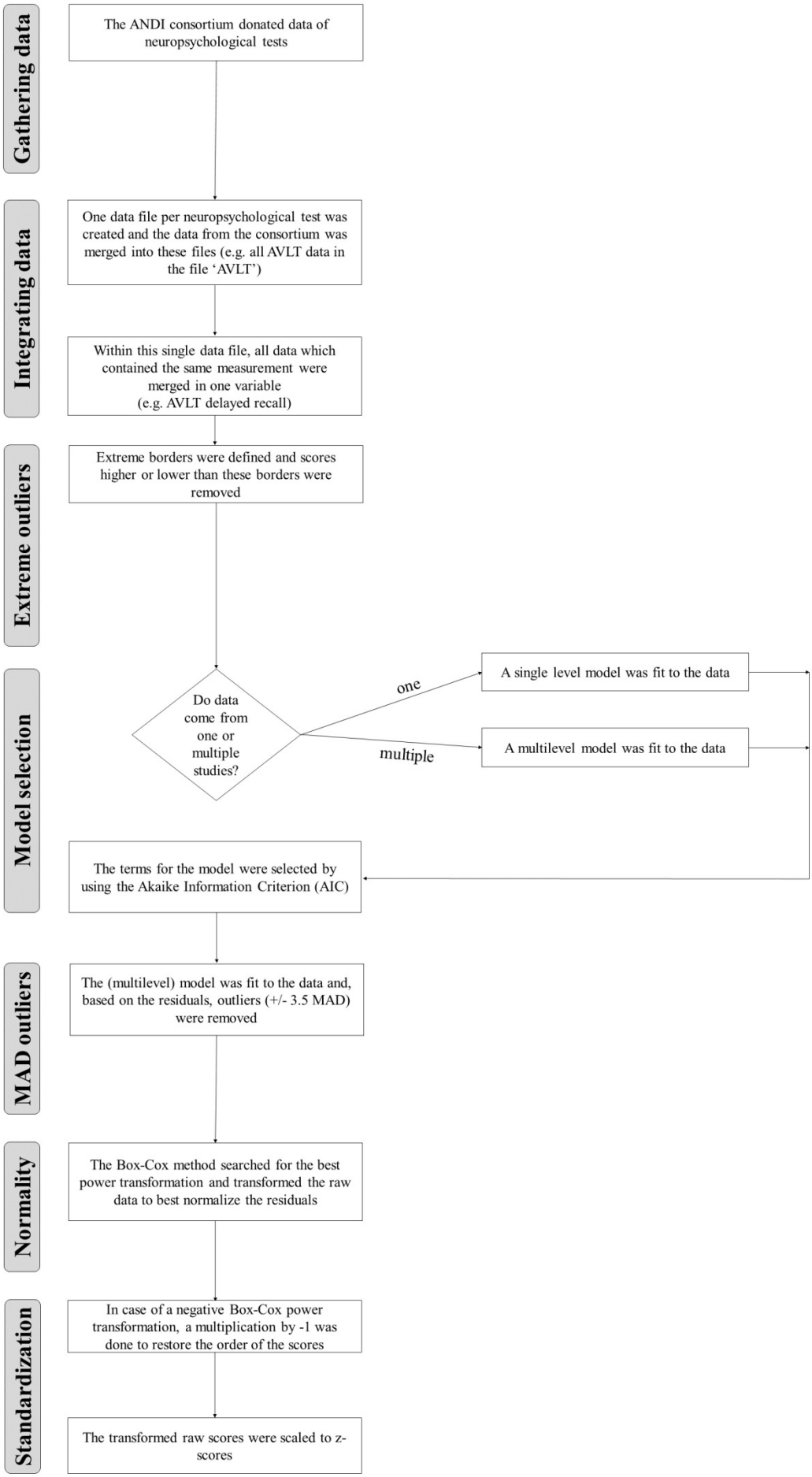


Figure 1. Flow chart describing all steps of the database construction.

2.1.1 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.

2.1.2 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.

2.1.3 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 be 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 three 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.

2.1.4 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 (Curran and 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 one stands for “no education” and seven 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.

2.1.4.1 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 five, 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.

Table 1: *Tabulation of number of participants by sex, age categories*, and level of education for the AVLT-delayed recall variable.*

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

**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.*

2.1.4.2 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 et al., 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).

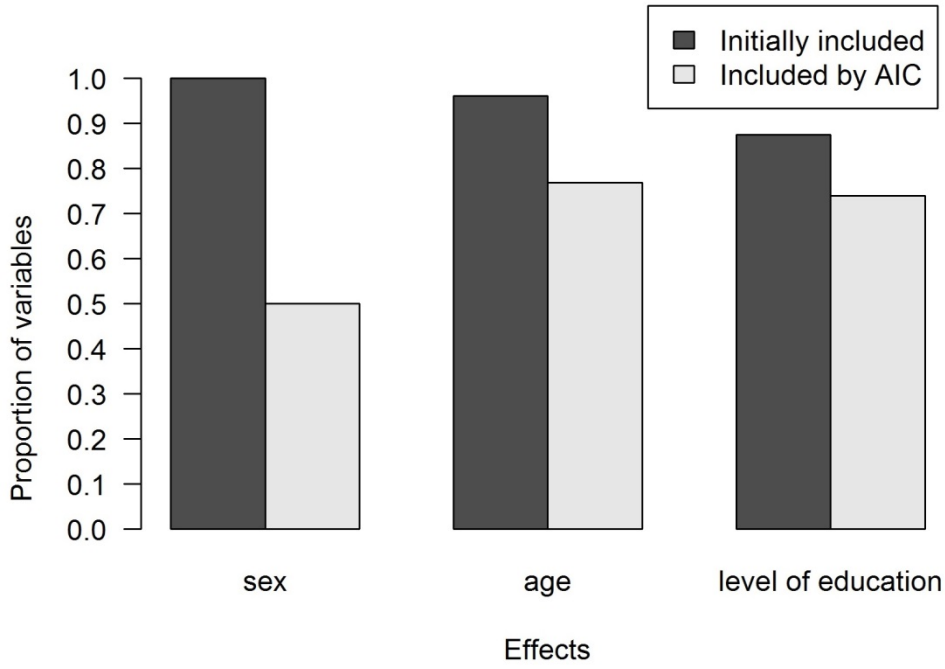


Figure 2. Proportion of variables for which the demographic effects were included in the models**.** In dark gray, effects that could be included after accounting for sample size constraints. In light gray, effects that were included after using the Akaike Information Criterion (AIC) to select effects.

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

2.1.5 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 deviation 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.

2.1.5.1 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.

2.2 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 and 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 and 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 Figures 3, 4, it can be seen that the residuals were somewhat skewed before transformation and were neatly normally distributed after transformation.

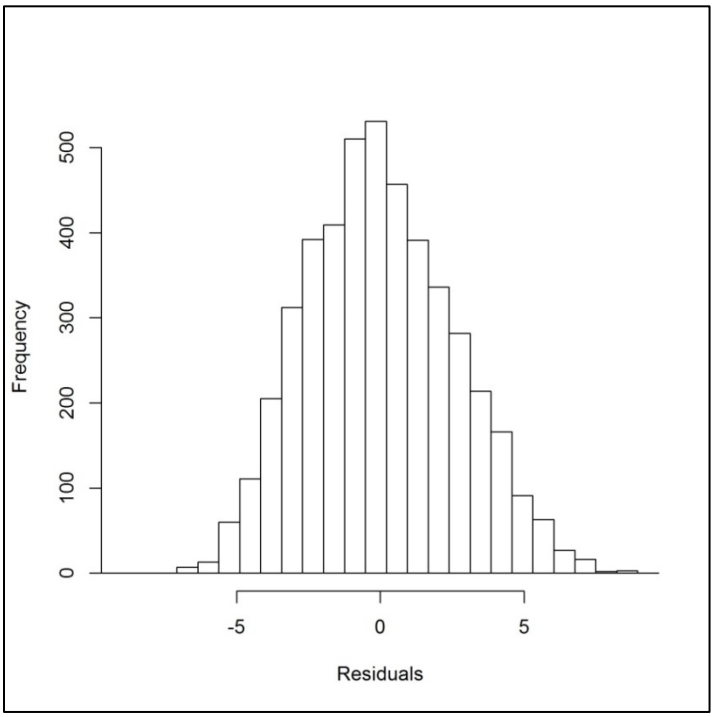


Figure 3. Distribution of the residuals of the model fitted to the AVLT delayed recall variable before power transformation.

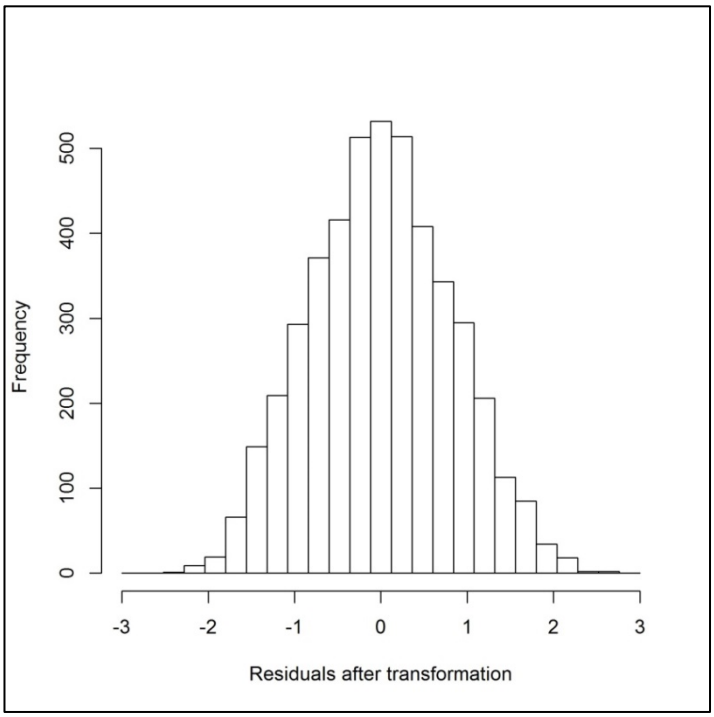


Figure 4. Distribution of the residuals of the model fitted to the AVLT delayed recall variable after power transformation.

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 Model Evaluation

2.3.1 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.

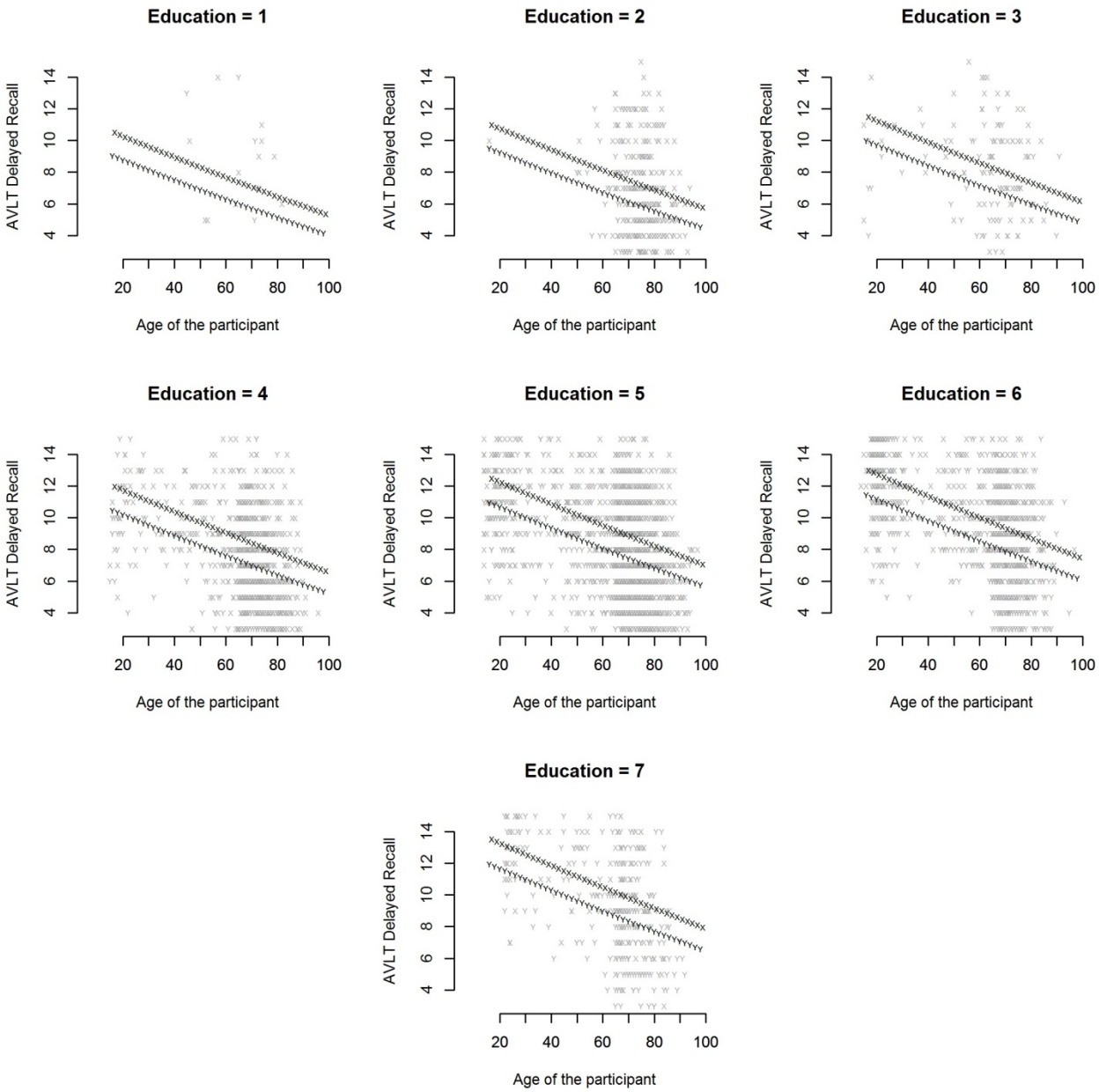


Figure 5: Raw scores on the AVLT delayed recall variable are plotted against age. Separate plots were made for the different levels of education. Men are depicted with the letter y and women with x.

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.

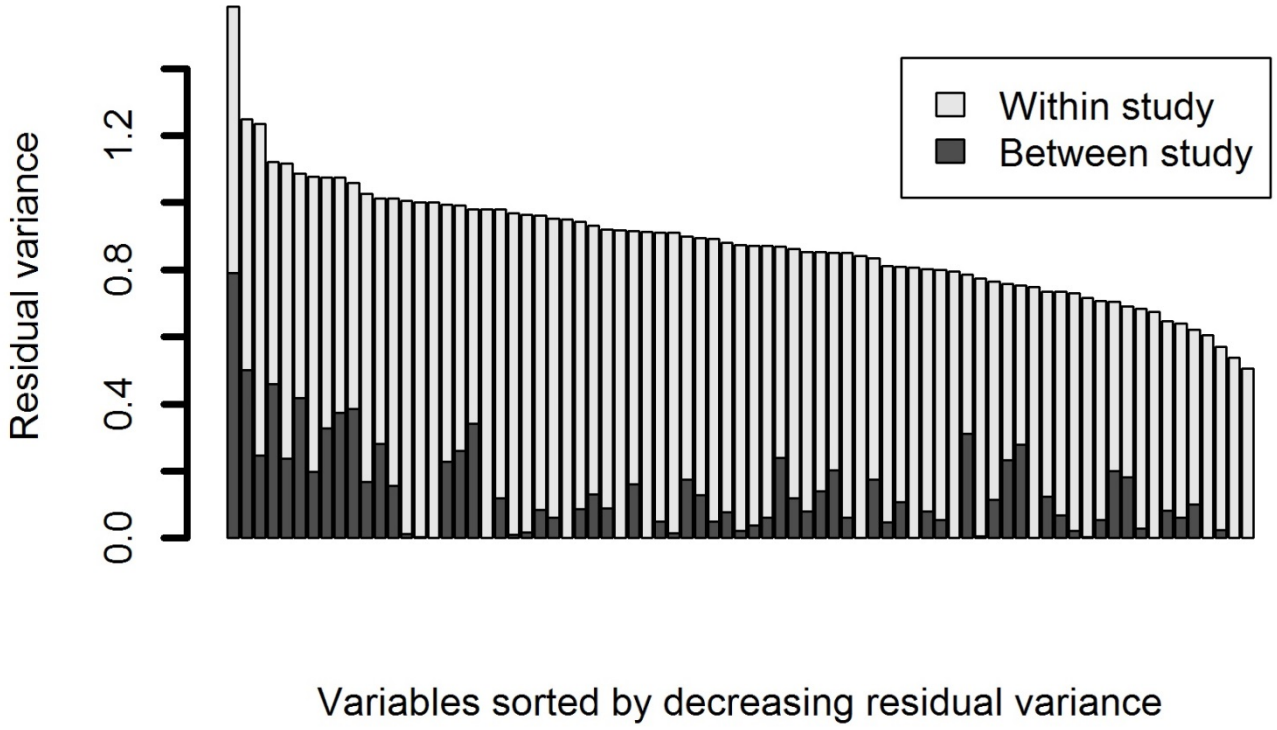


Figure 6: Partitioning of total residual variance for variables that were administered in multiple studies**.** Dark gray portions of the bars are the residual variance due to between study differences. Light gray portions of the bars are the residual variance due to within study/between participant differences.

2.4 Contents of ANDI

ANDI currently contains data of 26,635 healthy participants on 43 neuropsychological tests from different cognitive domains. As an example, Table 2 lists a selection of variables currently included in the database (the complete list is available on www.andi.nl).

Table 2. *Example variables per neuropsychological test. The number of participants, number of studies, and demographic information refer to one example variable.*

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 card in seconds)	30	2147	16-91	43	1-7
Stroop (CW interference in seconds)	30	2132	16-91	43	1-7
Visuospatial					
Judgement of Line Orientation	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 (Version 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

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.

2.4.1 Benefits of the ANDI database

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

2.4.1.1 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, Winicki, Pearlson, Gordon, & Schretlen, 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.

2.4.1.2 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 and Garthwaite, 2002), become feasible. With the ANDI database and the accompanying website, such comparisons can be routinely applied.

2.4.1.3 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.

2.4.2 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 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.

Chapter 3

Predicting Progression to Parkinson's Disease Dementia using Modern Neuropsychological Techniques.

This chapter has been published as:

J.A. Agelink van Rentergem*, N.R. de Vent*, H.M. Huizenga, J.M.J. Murre, ANDI Consortium, & B.A. Schmand, (2019). Predicting Progression to Parkinson's Disease Dementia using Modern Neuropsychological Techniques. *Journal of the International Neuropsychological Society*, 25(7), 678-687

*Shared first authorship

Abstract

Objective: Parkinson's disease with mild cognitive impairment (PD-MCI) is a risk factor for progression to 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 provides a more sensitive tool for predicting dementia status at three- and five-year follow up. This method allows a formal evaluation of a patient's profile of test scores given a large aggregated database with regression-based norms.

Method: The 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. The outcome variable was progression to dementia within three and five years.

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: We conclude that modern statistical techniques allow for a more sensitive prediction of PDD than the traditional PD-MCI criteria.

Many Parkinson 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., 2005). Mild Cognitive Impairment (PD-MCI) is predictive of progression to Parkinson 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 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 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 for a PD-MCI diagnosis. 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 comparing patients to such published norm tables. First, as the normative data for neuropsychological tests have been collected for each test separately, correlations between tests are usually unknown (except in case 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 (i.e. an abnormal score profile; Huizenga, Smeding, Grasman, & Schmand, 2007). Third, norm tables do not always allow for correction for the influence of demographic variables, even though age, sex, and level of education are known to influence the scores on neuropsychological tests. Moreover, it is often not possible to simultaneously correct for age, sex, and level

of education (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, Iverson, & Brooks, 2009).

In this study, we apply a new statistical method that circumvents these problems, to detect cognitive abnormality in newly diagnosed PD patients and 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 evaluates a patient's entire 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 a traditional, univariate normative comparison (Crawford & Garthwaite, 2002; Huizenga et al., 2007; Su et al., 2015). The database contains information about demographic variables and thus allows simultaneous correction for age, sex, and level of education. By using regression-based demographic corrections, drastic changes in the interpretation of test scores when moving from one to another norm table, are prevented. The new statistical method keeps the false positive rate under control, because it entails a single statistical comparison.

In this article, we will discuss 1) using an aggregate normative database instead of norm tables, and 2) using it to make multivariate normative comparisons. To examine whether these approaches are a good alternative to traditional (univariate) normative comparisons, we compare their performance to that of the PD-MCI criteria with traditional norms. We use existing data from a longitudinal study conducted by our group (Broeders et al., 2013). First, we compare the ability of the traditional PD-MCI criteria to predict PDD after three and five years to the same 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, we explored whether the new approach can give insight into which cognitive domains in particular are impaired in PD-MCI patients who progress to PDD compared to those who do not.

3.2 Method

3.2.1 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; Folstein, Folstein, & McHugh, 1975). After three years, the clinical status was missing for 26 patients. After five years, information was no longer available for another 24 patients (see Table 1). The institutional review boards of the participating hospitals approved the original study by Broeders et al. (2013) in accordance with the Helsinki Declaration.

Table 1. *Sample characteristics of the PD group from Broeders et al. (2013).*

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

3.2.2 PD-MCI

Broeders et al. (2013) applied the PD-MCI level II criteria (Litvan et al., 2012) as follows: 1) Patient has a PD diagnosis. 2) Gradual cognitive decline as reported by the patient or observed by the caregiver or the clinician, 3) 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. To compensate for the fact that gradual cognitive decline as observed by the caregiver or the clinician from criterion 2 was not available in our data, patients could also be diagnosed with PD-MCI if they reported no subjective complaints but had impairments (of at least 1.5 SD) on four or more tests. The reasoning behind this choice was that such broad impairments would surely be noticed by caregivers. With respect to the fourth criterion, patients were excluded if they had a score <24 on the Mini-Mental State Examination (Folstein et al., 1975).

3.2.3 PDD

PDD was used as the outcome variable. PDD at 3 and 5 years follow up was diagnosed by MDS criteria (Emre et al., 2007). Criteria were as follows: 1) a diagnosis of PD prior to the onset of dementia, 2) MMSE score lower than 24, 3) no depression, 4) cognitive deficits that are severe enough to interfere with daily functioning, as measured by the Behavioral Assessment of Daily Living (Collin, Wade, Davies, & Horne, 1988), Schwab & England Scale (Schwab & England, 1969), and Functional Independence Measure (Van Putten, Hornbart, Freeman, & Thompson, 1999), 5) an abnormal score on at least two of the following tests: clock drawing (Lezak et al., 2012), pentagon copying or serial 7s of the MMSE (Folstein et al., 1975).

3.2.4 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 (Nelson, 1976) as its score distribution was extremely skewed, violating the assumptions of the parametric normative comparisons that are used throughout this article. We replaced it by the Tower of London (Culbertson, & Zillmer, 1998) as an alternative test for the executive functions domain. An overview of the tests can be found in Table 2.

3.2.5 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=26,635) which consists of healthy individuals who either participated as control subjects in clinical studies, or participated in community-based studies. Since each participant completed only a subset of the tests that are included in ANDI, the number of participants per test varies between 62 and 5017 depending on the test. Table 2 gives an overview of the tests used for the present analyses and of the number of participants per test. All scores in the ANDI database have been transformed to normality and standardized to demographically corrected z-scores. For most test variables, age, sex, and level of education had a significant effect, and thus were included in the demographic correction (de Vent et al., 2016).

Table 2. *Characteristics of the neuropsychological test variables in ANDI.*

	N	% Men	Age range	Demographic variables[a]
Memory				
RAVLT - total (Rey, 1968).	5017	50%	18-97	A + S + E
RAVLT – delayed recall (Rey, 1968).	4540	49%	18-97	A + S + E
RBMT - Story subtest - immediate recall (Wilson, Cockburn, & Baddeley, 1983).	346	40%	19-90	A + S + E
RBMT - Story subtest - delayed recall (Wilson, Cockburn, & Baddeley, 1983).	353	40%	19-89	A + S + E
Language				
BNT – 30 item (Kaplan, Goodglass, & Weintraub, 1983).	467	42%	18-89	A + S + E
WAIS-III Similarities (Wechsler, 1997).	274	36%	18-80	E
Executive Functions				
COWAT (Benton & Hamsher, 1983).	2894	48%	18-97	A + S + E
TOL – total movement score (Culbertson & Zillmer).	62	53%	40-80	A
Visuospatial/constructive skills				
JOLO (Benton, Hamsher, Varney, & Spreen, 1983).	69	54%	40-80	S + E
Clock Drawing Test (Royall, Cordes, & Polk, 1998).	167	46%	40-82	E
Attention				
WAIS-R Digit Symbol Test (Wechsler, 1981).	2122	43%	18-91	A + S + E
TMT – part A (Reitan, 1992).	3216	47%	18-97	A + S + 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, A, S, and E indicate whether age, sex, and level of education were included for each variable. RAVLT = Rey Auditory Verbal Learning Test, RBMT = Rivermead Behavioral Memory Test, BNT = Boston Naming Test, WAIS-III = Wechsler Adult Intelligence Scale 3rd edition, COWAT = Controlled Oral Word Association Test, TOL= Tower of London, JOLO = Judgement of Line Orientation, WAIS-R = Wechsler Adult Intelligence Scale Revised edition, TMT = Trail Making Test.

3.2.5 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 with the usual way of working is that the ANDI data have been treated in a consistent manner across all tests (de Vent et al., 2016): For each test, the same procedures were followed for outlier removal, test score standardization, and selection of transformations to normality. Another difference is that for many tests a larger normative sample is available. Instead of z-values, ANDI uses t-values because these have better statistical properties (Crawford & Garthwaite, 2002). We interpreted these test statistics with the same statistical threshold for impairment as was used in the other PD-MCI analysis: z = -1.5 corresponds to a threshold p-value of 0.067 one-tailed. Because tests were one-tailed, only deviations in the negative direction were classified as impaired.

3.2.6 Abnormality as defined by MNC

Finally, used multivariate normative comparisons (MNC) (Huizenga, et al., 2007) to assess abnormal profiles. 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 age, sex, and level of education (Agelink van Rentergem et al., 2017a; Agelink van Rentergem et al., 2017b). The MNC procedure results in a p-value, which indicates abnormality when it is below a certain threshold. We tested for impairment (one-tailed), i.e. only deviations in the negative direction were classified as impaired. In univariate comparisons, if no subjective complaints were present, we required four instead of two significant results. In the MNC this adaptation is not possible, as only one test result is obtained. Therefore, we used different threshold values for those with and without subjective complaints, to determine whether results were significant. For patients with subjective complaints, we used a threshold p-value of 0.067*2 = 0.134. For patients without subjective complaints, we used a stricter threshold p-value of 0.067.

3.2.7 Analysis

We calculated whether the classification of cognitive impairment at baseline is predictive of progression to PDD. Sensitivity and specificity were compared across these three methods. Sensitivity was calculated by dividing the number of patients who were classified as impaired at baseline and progressed to 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 progress to PDD, by the total number of patients who did not develop PDD. This was done separately for the progression to PDD after three years, and after five years.

3.3 Results

3.3.1 Demographic characteristics

In Table 3, demographic and clinical characteristics of the patients are given, where the patients are separated into cognitively normal and abnormal categories using each of the three methods.

Table 3. *Demographic and clinical characteristics for the three groups (PD-MCI criteria, ANDI PD-MCI criteria and ANDI MNC) at baseline.*

	PD-MCI		ANDI PD-MCI		ANDI-MNC	
	Normal cognition	PD-MCI	Normal cognition	PD-MCI	Normal cognition	abnormal cognition
	N = 80 (65%)	N = 43 (35%)	N = 90 (73%)	N = 33 (27%)	N = 91 (74%)	N = 32 (26%)
Age	65.1 (10.6)	68.0 (10.1)	64.4 (10.7)	70.8 (8.3)	64.8 (10.1)	69.8 (10.7)
Sex M/F	43/37	23/20	46/44	20/13	48/43	18/14
MMSE	28.0 (1.9)	27.0 (2.0)	28.1 (1.9)	26.5 (1.9)	28.0 (1.8)	26.5 (2.1)
Disease duration (months)	18.3 (8.9)	20.1 (13.4)	18.0 (8.8)	21.3 (14.6)	18.2 (8.8)	20.8 (14.9)
LED	139.0 (142.6)	149.9 (139.3)	139.1 (143.5)	152.9 (135.4)	138.4 (145.2)	155.5 (129.7)
UPDRS	15.8 (7.8)	19.4 (7.8)	16.2 (8.2)	19.4 (7.0)	16.1 (7.8)	19.9 (7.8)
H&Y	1.6 (0.7)	2.1 (0.7)	1.7 (0.7)	2.1 (0.7)	1.7 (0.7)	2.1 (0.7)
HADS	8.5 (6.6)	13.5 (7.5)	9.4 (7.1)	12.7 (7.6)	9.4 (7.2)	12.8 (7.1)
SE-ADL	91.2 (5.8)	88.1 (7.9)	90.6 (6.9)	89.1 (6.3)	90.2 (7.0)	90.0 (6.2)
BADL	19.7 (0.7)	19.4 (1.5)	19.6 (1.2)	19.6 (0.8)	19.6 (1.2)	19.7 (0.7)
LED	139.0 (142.6)	149.9 (139.3)	139.1 (143.5)	152.9 (135.4)	138.4 (145.2)	155.5 (129.7)
UPDRS	15.8 (7.8)	19.4 (7.8)	16.2 (8.2)	19.4 (7.0)	16.1 (7.8)	19.9 (7.8)
H&Y	1.6 (0.7)	2.1 (0.7)	1.7 (0.7)	2.1 (0.7)	1.7 (0.7)	2.1 (0.7)
HADS	8.5 (6.6)	13.5 (7.5)	9.4 (7.1)	12.7 (7.6)	9.4 (7.2)	12.8 (7.1)

[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, A, S, and E indicate whether age, sex, and level of education were included for each variable.

3.3.2 Progression to PDD

Figure 1 gives an overview of progression to PDD for each method evaluated. According to the criteria used by Broeders et al. (2013), at baseline, 35% of the PD patients had PD-MCI. 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). 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 at baseline 27% of the patients had PD-MCI. 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 at baseline 26% PD patients were considered to be MNC abnormal versus 74% who were not. After 3 years, 25% of the MNC abnormal PD patients had progressed to PDD and 50% had not. Of the group who were not MNC abnormal, 1% had progressed to PDD and 79% had not. After 5 years, 38% of the MNC abnormal PD patients had progressed to PDD and 19% had not. Of the group who were not MNC abnormal, 5% patients nevertheless had progressed to PDD while 54% had not.

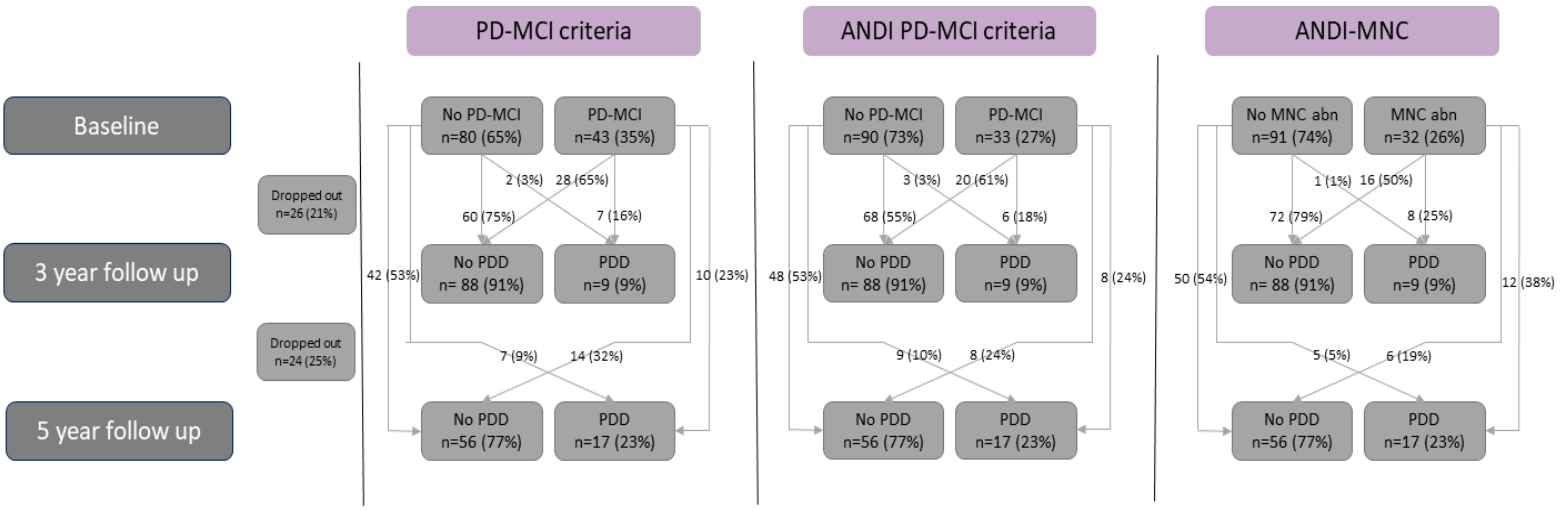


Figure 1. Progression of PD patients (n = 123) to PDD after 3 (n = 97) and 5 years (n = 73) for the three methods; PD-MCI criteria, PD-MCI criteria applied with ANDI, and multivariate normative comparisons (MNC) applied with ANDI.

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 abnormal could theoretically be different patients from those 33 classified as having PD-MCI using the ANDI method. The overlap in diagnoses between pairs of classification methods is explored in supplement 7.1.2 of chapter seven. Each of the three methods did indeed differ somewhat in the patients they classified as impaired, although the percentages of agreement were high (78-87%) and kappa's ranged from 0.49 to 0.68.

3.3.3 Sensitivity and Specificity

Sensitivity and specificity of the three methods are given in Table 4. After 5 years, sensitivity decreased for all methods but specificity remained similar or increased. However, the confidence intervals are large due to the small size of the samples that remained after attrition. Comparing the results of the original PD-MCI criteria to those obtained with the same criteria using ANDI shows a trade-off: sensitivity is higher for the original PD-MCI criteria, and specificity is higher for the PD-MCI criteria applied with ANDI. Therefore, there does not seem to be a clear advantage for either method.

Multivariate normative comparisons as applied with ANDI fare better than the other two methods. At both the three-year and the five-year follow-up, sensitivity and specificity were higher for the multivariate method than for the two univariate methods.

Table 4. *Sensitivity and specificity for progression to PDD of each method (original PD-MCI criteria, PD-MCI applied with ANDI, and MNC method applied with ANDI), specified for three- and five-year follow-up. In parentheses: 90% confidence interval (Agresti, & Coull; 1998).*

	Three-year follow-up		Five-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)

3.3.4 Cognitive Domains

We explored which cognitive domains were most often impaired in PD patients who were MNC abnormal, and whether there was a distinct profile for the patients who progressed to PDD. Figure 2 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 abnormal at baseline (solid lines), mainly showed impairment on the Rivermead Behavioural Memory Test and were slightly more impaired on the TMT part a and the WAIS-R Digit Symbol Coding task compared to those who were not MNC abnormal at baseline (dashed lines). Performance on the WAIS-R Digit Symbol Coding task seemed to be low for both those who were MNC normal and MNC abnormal at baseline. This is probably due to Parkinson pathology affecting motor performance.

No clear difference is visible between those who progressed to PDD after five years (red lines), and those who didn't (green lines). From the figure, we can see that the MNC abnormality is primarily driven by the AVLT, RBMT and letter fluency variables, as this is where the deepest troughs are for the MNC abnormal group (solid lines). If we look at just these tests, those who eventually do develop PDD (black solid line) are those with low scores on AVLT and letter fluency. Therefore, these tests seem to be most sensitive. The figure in supplement 7.1.1 of chapter seven plots a line for every individual patient, and thus provides more detailed information on individual differences.

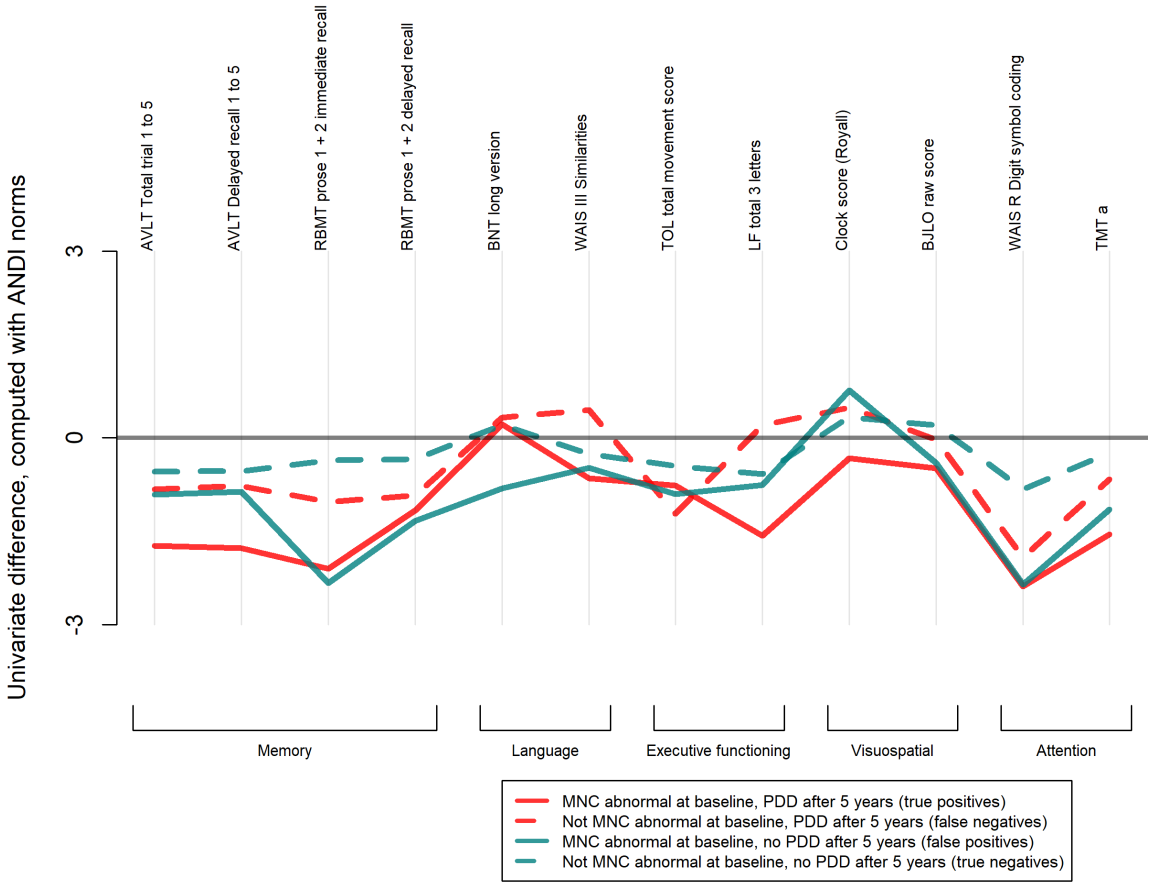


Figure 2. Mean demographically corrected z-scores for PD patients at baseline. Red indicates PD patients who progressed to PDD after five years. blue indicates no PDD after five years. The solid line is MNC abnormal, dashed is not MNC abnormal.

3.4 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. (2013) 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 to ours (1.5 SD deviating on at least two out of ten tests) show that between 21% and 60.5% of PD patients are diagnosed as 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 of progression to PDD, the MNC method applied with the ANDI database performed best. Sensitivity and specificity were higher for this method than for the other two PD-MCI criteria methods. This suggests that the improvement is mainly due to use of a multivariate statistical technique and not to use of a large aggregated database. This was true for both the prediction of PDD after three and after five years. 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 with Broeders et al. (2013). Just using the ANDI database instead of traditional norms therefore does not seem to improve prediction.

Figure 3 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 3, for all three methods a decrease in sensitivity can be observed between the three-year and five-year follow-up. An explanation would be that with a short period between baseline and PDD diagnosis, most patients who progressed to dementia were already rather severely impaired, leading to a high sensitivity. With more time between baseline and PDD diagnosis, some patients who progressed to dementia may have been unimpaired at baseline, leading to a lower sensitivity. Similarly, a small increase in specificity between the three-year and five-year follow-up can be observed. This is explained by the time it takes to progressed to 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 progressed to dementia, leading to an increase in specificity.

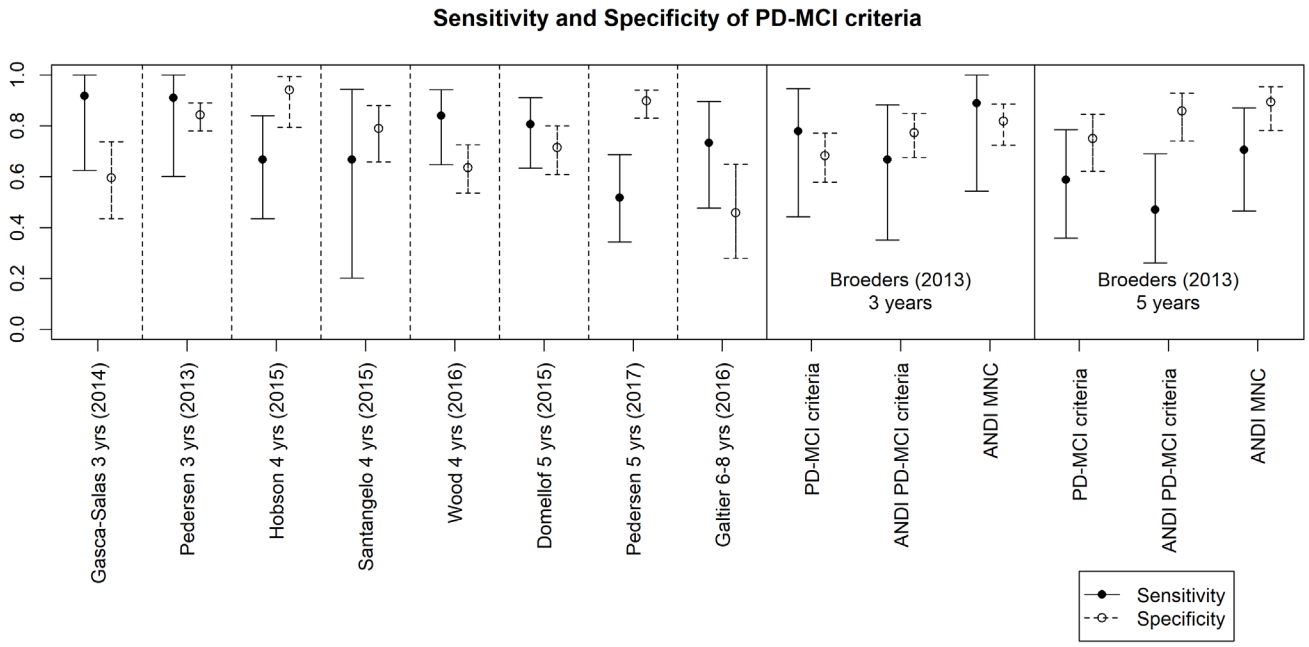


Figure 3. Sensitivity and specificity of the PD-MCI criteria for progression to PDD when using 1.5 SD as a cut off score in various previous studies (left panels), and sensitivity and specificity of the three methods investigated in the current study (right panels). Error bars indicate 95% confidence intervals (Agresti & Coull, 1998).

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 the tables in supplement 7.1.2 of chapter seven a specification of which patients were lost to follow-up is given). Because a formal test of a difference between rates requires a very large sample size if the prevalence of a disease is low (Carley, Dosman, Jones, & Harrison 2005), the power to detect differences between the sensitivity and specificity of the different methods was low, and confidence intervals were broad. We recommend that future studies either collect a much larger sample, or find a way to synthesize the literature to obtain a better estimate of these rates.

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, 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 stricter criterion was used for MNC.

It is not easy to apply multivariate normative comparisons in daily clinical practice. We therefore developed the user-friendly ANDI website. Currently, ANDI is only applicable to the Dutch-speaking population. However, we provide it as an open source infrastructure. It is possible to copy ANDI and re-create the system in other countries, which would only require a local (aggregated) normative sample. Doing so would make multivariate normative comparisons easier applicable in the clinical setting.

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.

Chapter 4

An Operational Definition Of 'Abnormal Cognition' To Optimally Predict Progression To Dementia.

What are optimal cut-off points for univariate and multivariate normative comparisons?

This chapter is published as:

N.R. de Vent, H.M. Huizenga, J.M.J. Murre, J.A. Agelink van Rentergem, W.M. van der Flier, S.A.M., Sikkes,K van der Bosch, ANDI Consortium, & B.A. Schmand (2020). An Operational Definition Of 'Abnormal Cognition' To Optimally Predict Progression To Dementia. What are optimal cut-off points for univariate and multivariate normative comparisons? *Journal of Alzheimers's Disease, 1-11*

Abstract

Objective:

In clinical neuropsychology and neurology, there is no consensus on a single definition of abnormal cognition. The aim of this study was two-fold. First, we examined how to operationally define 'abnormal cognition' in order to optimally predict progression to dementia in a memory clinic sample. Second, we tested whether a profile analysis of cognitive test results would improve the prediction of progression to dementia compared to standard clinical evaluation.

Method: We used longitudinal data from 835 non-demented patients (mean age 64.3 (sd 8.5), median MMSE 28 (IQR 2)) of the Amsterdam Dementia Cohort. At follow-up after a median of 28 months, 182 (22%) patients had progressed to dementia. For 10 cognitive measures at baseline, we determined which combination of number of abnormal tests and magnitude of score deviations best predicted progression to dementia. For the multivariate profile analysis, we determined the magnitude of the abnormality of the profile that best predicted progression to dementia. In order to do so, we used a cross-validation method in which optimal criteria derived from a survival analysis of the first half of the sample were used to predict progression to dementia in a survival analysis of the second half.

Results: The predictive ability for progression to dementia of one, two and three abnormal test scores out of 10 is highly similar (Cox hazard ratios: 3.7 – 4.1) provided that cut-off values are adapted appropriately. That is, the cut-off value has to be less stringent if the number of abnormal tests required increases: the optimal cut-off is $z < -1.45$ when one deviating score is required, $z < -1.15$ when two abnormal tests are required, and $z < -0.70$ when three abnormal tests are required. The multivariate profile analysis has similar predictive ability when using the cut-off of $p < 0.22$ (hazard ratio 3.8). A likelihood ratio test showed that the profile analysis improves the prediction of progression to dementia when added to the standard clinical evaluation ($\chi^2(1) = 13.45$, $p<.001$).

Conclusion: Abnormal cognition may be defined as having one, two or three abnormal test scores out of 10 as long as the magnitude of the score deviations is adapted accordingly. An abnormal multivariate profile of test scores predicts decline to dementia equally well to the traditional univariate approach. Also, the multivariate profile test increases the ability to predict progression to dementia when used complimentary to the standard clinical evaluation.

In clinical neuropsychology and neurology, there is no consensus on a single definition of abnormal cognition. Many definitions have been coined, such as Benign Senescent Forgetfulness (Kral, 1962), Age-Associated Memory Impairment (Crook et al., 1986), Cognitive Impairment, no dementia (Graham et al., 1997), Age-Associated Cognitive Decline (Levy, 1994), and Subtle Cognitive Decline (Thomas, Edmonds, Eppig, Salmon, & Bondi, 2018). Today, Mild Cognitive Impairment (MCI) is probably the widest known concept (Flicker, Ferris, & Reisberg, 1991; Petersen et al. 1999). In the DSM-5 this corresponds to Minor Neurocognitive Disorder (American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, 5th Edition.)

Apart from this lack of consensus on how to conceptually define abnormal cognition, an operational definition of abnormal cognition is complicated by some technical aspects. First, there is no consensus on the minimally required deviation of cognitive test scores (i.e., how many standard deviations [SD] below the mean). Typically, in the context of MCI, scores of -1.5 SD below the appropriate normative mean are considered abnormal, but some authoritative definitions consider scores below -1 SD abnormal (DSM-5, Albert, et al. 2011, Petersen & Morris, 2005). In the broader context of clinical (neuro) psychology, scores either in the fifth percentile (corresponding to- 1.65 SD) and below, or in the second percentile

(corresponding to -2 SD) and below, are thought to reflect impairment (Strauss, Sherman, & Spreen, 2006; Lezak, Howieson, Bigler, & Tranel, 2012). Second, there is no consensus on how many cognitive tests, or which proportion of the tests administered, should be abnormal to meet criteria for impairment. Moreover, the required number and magnitude of score deviations, are probably not independent. When one requires more deviating scores in order to classify a person as impaired, the magnitude of these score deviations can probably be smaller than when one requires only one abnormal score. This also depends on the number of tests administered.

In addition to the lack of consensus on conceptually and operationally defining abnormal cognition, there are additional issues that may complicate the interpretation of the neuropsychological examination. First, to evaluate test results, neuropsychologists generally use normative data published in test manuals. Each test has its own normative sample, which may vary in quality and the norms may not always be optimal for each age group. For example, norms for oldest patients are often lacking, some norms were collected decades ago, norms may not always be available in the patient's language. Second, demographic variables such as age, sex, and level of education influence scores on cognitive tests and should be considered (Lezak et al., 2012). Not all published norms, however, allow correction for each of these variables. Third, because normative data are mostly collected for each test separately, correlations between tests are not taken into account. Thus, one cannot jointly evaluate scores of several tests. It is therefore impossible to formally judge whether score combinations are unusual. Often clinicians have an intuitive feeling for abnormalities in the cognitive profile, but formal evaluation of whether particular profiles of scores are normal or abnormal is impossible.

The Advanced Neuropsychological Diagnostics Infrastructure (ANDI, www.andi.nl) (de Vent, Agelink van Rentergem, Schmand, Murre, & Huizenga, 2016) introduces a new way of evaluating neuropsychological test results that solves many of these problems. ANDI consists of a large, representative, aggregated database with neuropsychological test data from over 24,000 healthy people. ANDI provides regression-based norms that are simultaneously corrected for all relevant demographic variables (Testa, et al. 2009). Moreover, ANDI has information on correlations between tests and therefore provides the possibility to

evaluate the *profile* of test scores by multivariate normative comparison (MNC). This method is a more sensitive way to analyze neuropsychological tests results (Huizenga, Smeding, Grasman, & Schmand, 2007; Agelink van Rentergem, Murre, & Huizenga, 2017, Agelink van Rentergem, de Vent, Schmand, Murre, & Huizenga, 2018). Using such a profile analysis enables the detection of subtle cognitive deficits that would go unnoticed in traditional analysis of neuropsychological test results (van Rentergem, de Vent, Huizenga Murre, & Schmand 2016; Su et al, 2012).

The first aim of the current study was to find the optimal combination of magnitude and number of abnormal scores in order to operationalize the definition of abnormal cognition. Our criterium was success to predict progression to dementia. That is, we examined which number of abnormal tests and which magnitude of score deviations at baseline best predicted progression to dementia. In doing so, we also included often used solutions, such as one abnormal score of 1.5 SD below the appropriate mean. Secondly, we investigated what cut-off score should be used for the multivariate profile analysis offered by ANDI in order to best predict progression to dementia. After the optimal cut-off scores had been determined, we investigated how well they predicted progression to dementia. Finally, we investigated whether the multivariate profile analysis can be used in standard clinical practice to better predict progression to dementia. In order to address these aims, we used neuropsychological data from memory clinic patients and normative data from the ANDI database.

4.2 Method

4.2.1 Patients

The patient data came from the Amsterdam Dementia Cohort (van der Flier & Scheltens, 2018). The patients were referred by general practitioners and other health care professionals to Alzheimer center Amsterdam, Amsterdam UMC because of cognitive complaints. All patients had their baseline visit at this memory clinic between January 1993 and March 2016. In total, 1004 non-demented patients were considered for inclusion. Patients were excluded if they 1) had received a dementia diagnosis at baseline, 2) had not been followed-up, or 3) had more than one missing value on the cognitive test battery (see below). This left 835

(for an overview see Figure 1). The mean age was 64.3 years (SD=8.5, range 32 – 87) and 61% of the patients were male. The follow-up period ranged from 6 to 214 months (17 years) (median number of months was 28). The local ethical review board approved the study, and all patients provided informed consent for their clinical data to be used for research purposes.

4.2.1.1 Diagnostic procedures.

The diagnostic work-up has been described in detail elsewhere (van der Flier & Scheltens, 2018). In brief, for each patient it consisted of a medical and neurological evaluation by a neurologist, an assessment of vital functions, informant-based history, assessment of needs by a specialized dementia nurse, a neuropsychological examination, magnetic resonance imaging, and ancillary laboratory assessments. Each year, the patients were invited for a follow-up visit with a neurologist and a neuropsychologist. The diagnosis at baseline and at each follow-up visit was evaluated in a multidisciplinary consensus meeting.

Patients were labeled as MCI (N=420), if they fulfilled the Petersen criteria (Petersen, et al. 1999) until 2012 and the NIA-AA criteria (Albert, et al. 2011) from 2012 onwards. Patients were labeled as having subjective cognitive decline (SCD) (N=415), when clinical investigations were normal and cognitive performance was within normal limits (i.e., criteria for MCI, dementia or psychiatric disorder not fulfilled, Jessen et al., 2014). These decisions were based on test score evaluations using conventional, published normative data, and on all other available information.

4.2.1.2 Dementia diagnosis.

At follow-up, dementia due to AD was diagnosed according to the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association criteria (until 2012) or NIA-AA criteria (Dubois, et al. 2007; McKahnn, et al. 2011). Other forms of dementia were diagnosed according to consensus criteria (Román, et al. 1993; Neary, et al. 1998; McKeith, et al. 2005; Rascovsky, et al. 2011). For the current study, all forms of dementia were collapsed, so the main outcome measure was any type of dementia.

4.2.2 Materials

4.2.2.1 Neuropsychological tests.

All neuropsychological tests that overlapped between the patient test battery and the ANDI database were used for the current analysis. As a screening tool for cognitive impairment, the Mini Mental State Examination (MMSE) was used (Folstein, Folstein & McHugh, 1975). To assess memory, the Dutch version of the Rey Auditory Verbal Learning Test (Rey, 1964) immediate and delayed recall was used. To assess verbal fluency, category fluency (animals) and the Dutch version of the Controlled Word Association Test (letter fluency) were used (Benton & Hamsher, 1983). For attention, the Trail Making Test part A (Reitan, 1992) and Stroop color word test (Stroop, 1935) color naming and word reading were used. For executive functioning, the Trail Making Test part B (Reitan, 1992) and the Stroop color word test (Stroop, 1935) interference condition were used. These six tests (10 test variables) were used in the current analysis. Note, however, that the total administered battery varied over the years and was flexible to some extent. That is, some patients received additional tests, and some received fewer or a different selection of tests, depending on what the neuropsychologist deemed necessary. In general, more tests were administered in the clinical assessment than were used in the current analysis.

ANDI normative control sample and profile of test scores

All scores in the ANDI database have been transformed to normality and standardized to demographically corrected *z*-scores, the matching corrections were applied to patients' scores. A detailed overview of the ANDI data handling procedure can be found in de Vent *et al.* (2016).

Supplement 1 gives an overview of tests used in the current analyses and of the characteristics of the ANDI normative data.

4.2.3 Analyses

Univariate: *Calibration of number and magnitude of score deviations*

The first step was to operationally define 'abnormal cognition'. We did this by determining both the number of tests that should be abnormal and the magnitude of the abnormality in order to optimize prediction of dementia at follow-up. We randomly split our data in two sets. In the first set (N=416; *calibration set*), we selected the number and magnitude of deviations that best predicted dementia status at follow-up. In order to do so, we fitted Cox-regressions (proportional hazards regressions) (Cox, 1972) in which time from baseline to the event —diagnosis of dementia— was the dependent variable; dementia-free survival time was computed as number of days from baseline to the last follow-up without dementia. Age was included as a continuous covariate as the prevalence of dementia increases with age (Ott, Breteler, Harskamp, Stijnen, & Hofman, 1998). For each Cox-regression we varied the number of deviating test scores from 1 to 3 deviations (out of the 10 possible score deviations) and we varied the magnitude of the deviations in z-scores ranging from 0 to -3 (in 60 steps of 0.05). Thus, a total of 180 (3x60) combinations were evaluated. The combination with the best fit of the Cox-regression was the combination that best predicted dementia status at follow-up. Fit of the Cox-regression was evaluated using the concordance of the regression model (Harrell, Lee, & Mark, 1996). The concordance (or c index) is a measure of predictive discrimination and is defined as the proportion of all patient pairs in which the predictions (patient will decline to dementia or not) and outcomes are concordant. A value of 0.5 indicates no predictive discrimination and a value of 1.0 indicates perfect separation of patients with different outcomes.

Multivariate: *Calibration of magnitude of profile deviation*

We also calibrated the multivariate normative comparison (MNC) (Huizenga *et al.* 2007) by fitting Cox-regressions. Thus, we determined the magnitude of profile deviation best predicting progression to dementia. As the MNC-method gives a single test statistic that indicates whether or not a profile is abnormal, only the magnitude of the abnormal profile needed to be calibrated. An abnormal profile can be either positive (abnormally good) or negative (abnormally poor). Only the patients with a negatively

deviating profile were considered as having abnormal cognition. This amounts to one-tailed statistical testing.

For the univariate analysis of number and magnitude of score deviations, deviation was defined per variable. For the current multivariate analysis, deviation is defined for all variables at the same time. Therefore, we took the p-value of negatively deviating profiles as the measure of abnormality and varied the p-value cut-off to discover what p-value leads to best prediction of progression to dementia. We examined cut-off p-values ranging from 0 to 1 (in 1000 steps of 0.001) to select the p-value that best predicted which patients would progress to dementia.

Predicting dementia status at follow up

After the best cut-off scores had been calibrated, the second half of the data (N=419; *test set*) was used to test how well the univariate cognitive status (based on one test, two tests and three tests) and MNC-status predicted progression to dementia. Cox-regressions were conducted again with time from baseline to the event (diagnosis of dementia) as the dependent variable. The discrete independent variable was univariate cognitive status (based on one test, two tests and three tests) or the MNC status. Age was again included as a continuous covariate.

Diagnostic overlap

We also calculated Cohen's kappa's (Cohen, 1960) in order to investigate whether each method labeled the same patients as having abnormal cognition (inter-rater reliability).

Adding profile analysis to standard clinical evaluation

Finally, we tested whether the MNC profile analysis could be used complimentary to standard clinical evaluation. We performed Cox-regression in a stepwise procedure. In the first step, the syndrome diagnosis (MCI or SCD) as acquired in standard clinical practice was used to predict time from baseline to the event (dementia). Age was included as a continuous covariate. In the second step, the patients' MNC-status (normal or abnormal)

was also included to see whether it had added value over the MCI status to predict progression to dementia. Added value was tested by means of a likelihood ratio test.

4.3. Results

4.3.1 Progression to dementia

Over time, 182 of the 835 patients progressed to dementia. Of these, 133 had probable AD, five had possible AD, 10 had fronto-temporal dementia, 20 had vascular dementia, six had dementia with Lewy bodies, and eight had a different type of dementia. Figure 1 gives an overview of the inclusion of patients.

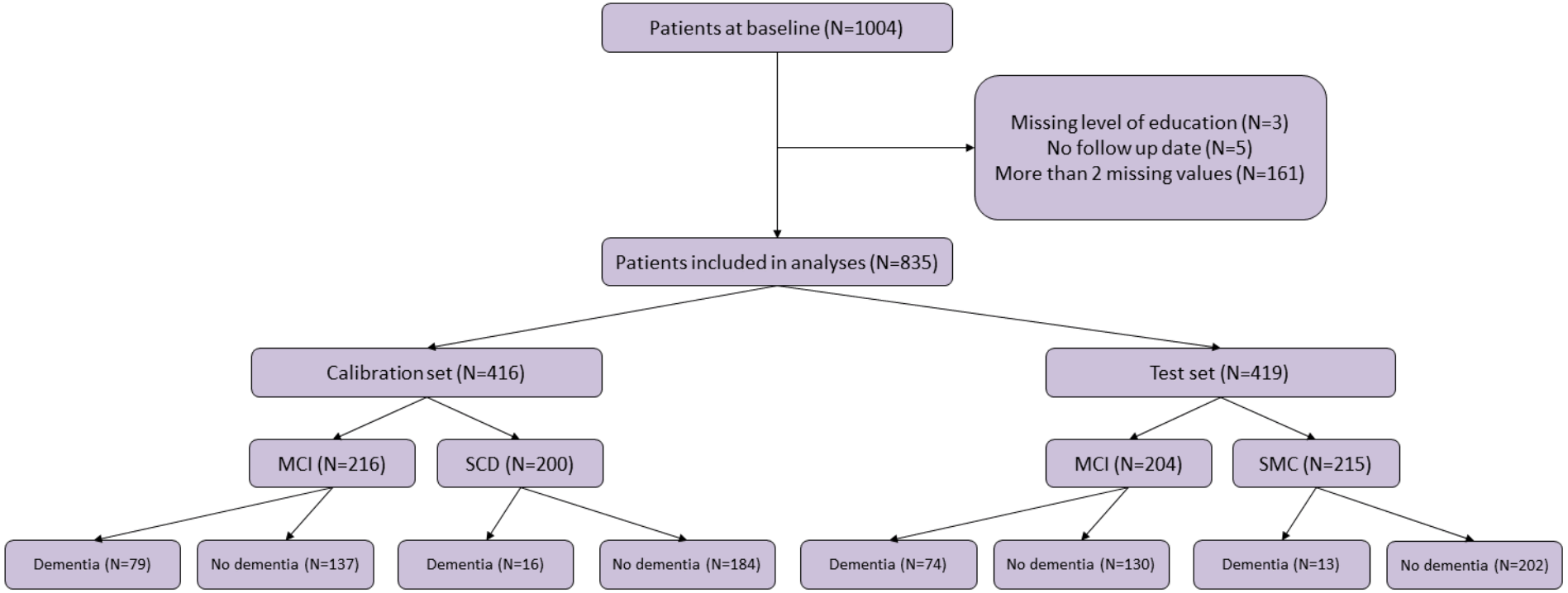


Figure 1. Flow chart describing the inclusion of patients and their progress to dementia over time.

4.3.2 Calibration of magnitude and number of test score deviations.

The first step was to investigate how many abnormal scores there should be and how large they should be to best predict progression to dementia. For this calibration, a random half (*calibration set*) of the data was used (N=416) (see supplementary online materials). One, two and three abnormal scores all produced similar results in terms of how well they predicted decline to dementia at follow-up, as the highest concordance obtained in all three was around 0.70. However, the cut off points producing the best result differed. For one abnormal test (z=-1.45) we needed to be stricter than for two (z=-1.15) and three abnormal scores (z=-0.70), and for two we needed to be stricter than for three.

For the MNC status, the calibration showed a *p*-value of 0.22 (one-tailed) had to be selected as the best value to predict progression to dementia. Thus, patients were predicted to progress to dementia if they showed an abnormal test profile worse than that of the lowest scoring 22% of the healthy population.

4.3.3 Testing the prediction of progression to dementia.

The second half of the data (the *test set*) (N=419) was used to assign patients into cognitively normal and cognitively abnormal based on the univariate operationalizations of cognitively abnormal (one, two and three abnormal scores) and MNC status. Table 1 gives an overview of characteristics of subgroups determined according to each method. Figure 2 gives an overview of progression to dementia for each method.

Table 1. *Demographic and clinical characteristics of the three groups, MCI (established by standard clinical evaluation), univariate cognitive status based on 1, 2 or 3 abnormal scores, or by MNC status.*

	Standard clinical evaluation		1 abnormal score z < 1.45		2 abnormal scores z < -1.15		3 abnormal scores z < - 0.70		MNC status	
	Normal cognition	MCI	Normal cognition	Abnormal cognition	Normal MNC-status	Abnormal MNC- status	Normal cognition	Abnormal cognition	Normal cognition	Abnormal cognition
	N=215 (51.3%)	N=204 (48.7%)	N=152 (36.3%)	N=267 (62.7%)	N=224 (53.5%)	N=195 (46.5%)	N=185 (44.2%)	N=234 (55.8%)	N=145 (34.6%)	N=274 (65.4)
Age (sd)	62.4 (8.8)	67.1 (7.1)	63.9 (7.7)	65.1 (8.7)	64.4 (7.9)	65.0 (8.8)	63.9 (8.0)	65.3 (8.6)	64.0 (7.8)	65.0 (8.6)
Sex M/F	130/85	140/64	96/56	174/93	149/75	121/74	123/62	147/87	89/56	181/93
MMSE	28.5 (1.3)	27.1 (1.7)	28.3 (1.2)	27.4 (1.8)	28.3 (1.3)	27.2 (1.8)	28.4 (1.3)	27.3 (1.8)	28.6 (1.3)	27.4 (1.7)
RAVLT – tot	40.4 (8.9)	31.0 (7.3)	41.9 (8.5)	32.3 (8.1)	39.3 (9.3)	31.8 (7.9)	40.8 (8.8)	31.9 (8.0)	41.9 (9.0)	32.6 (8.0)
RAVLT –delayed	8.1 (2.9)	3.8 (2.7)	8.6 (2.6)	4.5 (3.1)	7.7 (3.0)	4.1 (3.1)	8.0 (3.1)	4.4 (3.0)	8.3 (3.2)	4.8 (3.1)
Animal fluency	22.2 (5.4)	18.6 (4.7)	23.4 (5.0)	18.7 (4.8)	22.2 (5.3)	18.4 (4.6)	23.3 (5.1)	18.1 (4.4)	23.9 (9.7)	18.6 (4.7)
COWAT	35.8 (11.0)	32.9 (11.6)	39.5 (9.8)	31.9 (11.5)	36.6 (10.5)	31.7 (11.8)	38.8 (10.4)	30.8 (10.9)	39.4 (9.7)	31.7 (11.3)
TMTa	39.1 (13.9)	46.0 (17.5)	35.6 (10.7)	46.4 (17.4)	37.5 (11.2)	48.2 (18.8)	35.9 (10.5)	47.6 (17.8)	34.9 (9.8)	46.4 (17.3)
TMTb	89.6 (35.2)	129.6 (64.6)	83.9 (26.2)	123.4 (62.1)	89.6 (31.4)	131.7 (67.3)	85.2 (26.5)	127.8 (64.3)	83.6 (30.3)	122.5 (60.6)
Stroop- w	47.2 (11.9)	48.5 (10.5)	43.6 (5.6)	50.2 (12.9)	44.9 (6.1)	51.2 (14.1)	43.5 (5.5)	51.3 (13.3)	42.3 (4.6)	50.8 (12.6)
Stroop- c	64.4 (15.1)	68.0 (15.5)	59.2 (8.5)	70.2 (17.0)	60.9 (9.2)	72.3 (18.5)	59.3 (8.3)	71.7 (17.4)	57.6 (8.2)	70.7 (16.4)
Stroop- c w	108.5(29.2)	128.5 (41.2)	100.0 (18.4)	128.7 (41.4)	103.0 (21.5)	135.8 (43.7)	100.9 (21.7)	132.0 (41.4)	98.3 (21.8)	128.8 (39.6)

Abbreviations: MMSE=Mini-Mental State Examination; RAVLT=Rey Auditory Verbal Learning Test; COWAT=Controlled Oral Word Association Test; TMT=Trail Making Test; Stroop= Stroop Color-Word Test.

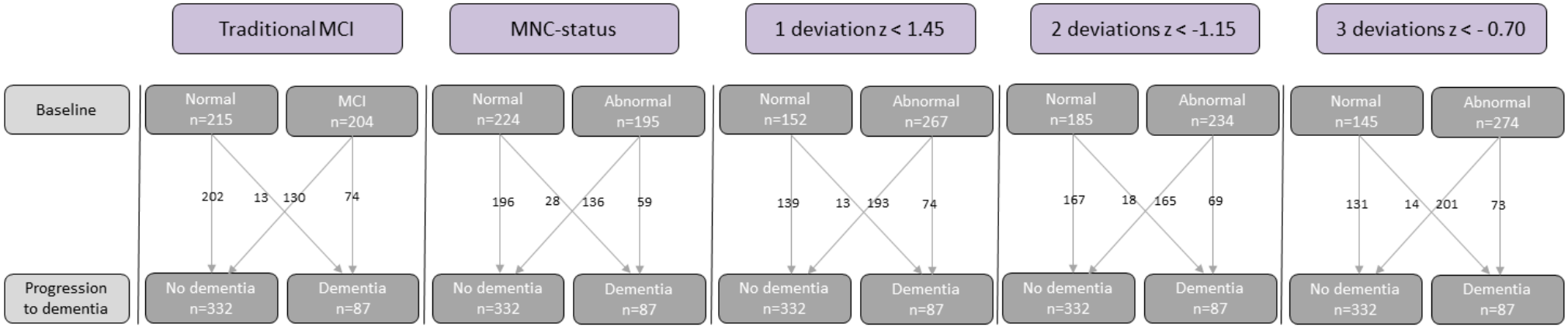


Figure 2: Classification of cognition by the five indicators of abnormal cognition: Conventional-MCI diagnosis, univariate cognitive status based on 1, 2 or 3 abnormal scores, and MNC status.

To test how well the univariate cognitive status (one, two and three abnormal scores) and MNC status predicted progression to dementia, the same Cox-regressions (with age as a covariate) were conducted. An overview of the results can be found in Table 2.

Table 2:
Hazard ratios, their confidence intervals, c-index and p-values of the survival analysis. The models have been corrected for age.

Factor	Hazard Ratio	95% CI	c-index	*p*
1 abnormal score, cut-off z < -1.45	4.13	2.28 – 7.47	0.70	<.001
2 abnormal score, cut-off z < -1.15	3.68	2.17 – 6.24	0.70	<.001
3 abnormal score, cut-off z < -0.70	4.02	2.26 – 7.16	0.73	<.001
MNC status, cut-off p < 0.22	3.77	2.39 – 5.97	0.72	<.001

4.3.3.1 Univariate cognitive status.

Because the 1, 2, and 3 abnormal scores predicted progression to dementia about equally well in the calibration process, we evaluated all three. Using one abnormal score of z < - 1.45, two abnormal scores of z <-1.15, or three abnormal scores of z < -0.70, showed that patients with abnormal cognition were about four times as likely to have a dementia diagnosis at follow up compared to patients with normal cognition. Figure 3 shows the survival curves for each of the three univariate methods.

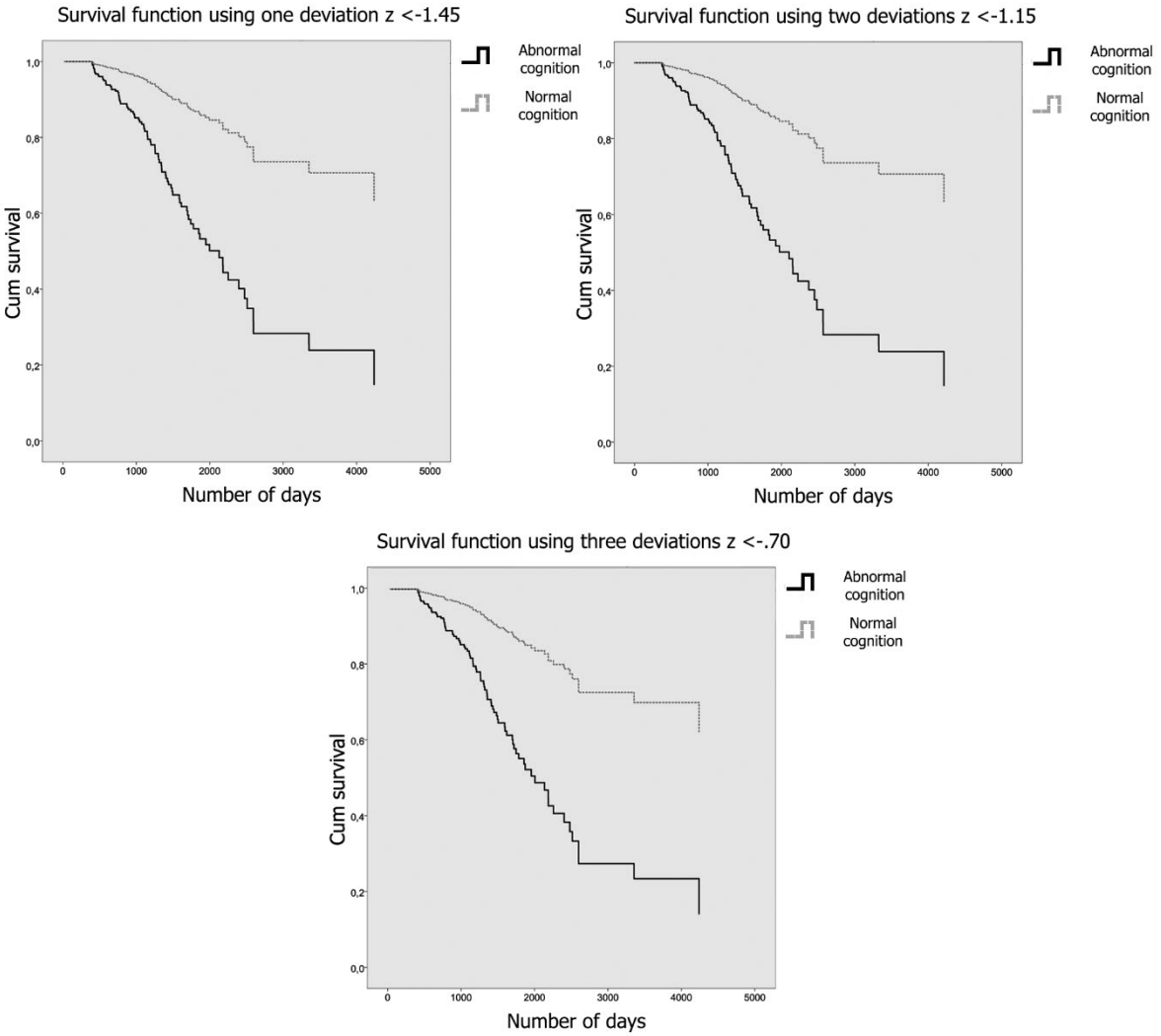


Figure 3. On the top left, the survival curves using one abnormal score of z < -1.45 are plotted. On the top right, the survival curves using two abnormal scores of z < - 1.15. Below, the survival curves using three abnormal scores of z < -.70. Patients with normal cognition are shown in dashed grey lines and patients with abnormal cognition in solid black lines. The y-axis shows cumulative (dementia free) survival across time, number of days participated is plotted on the x-axis.

4.3.3.2 MNC status.

The MNC profile analysis showed that when using the optimum cut-off score of p < .22, patients with an abnormal profile were almost 4 times as likely to have a dementia status at follow-up. Figure 4 shows the survival curves plotted for the MNC status.

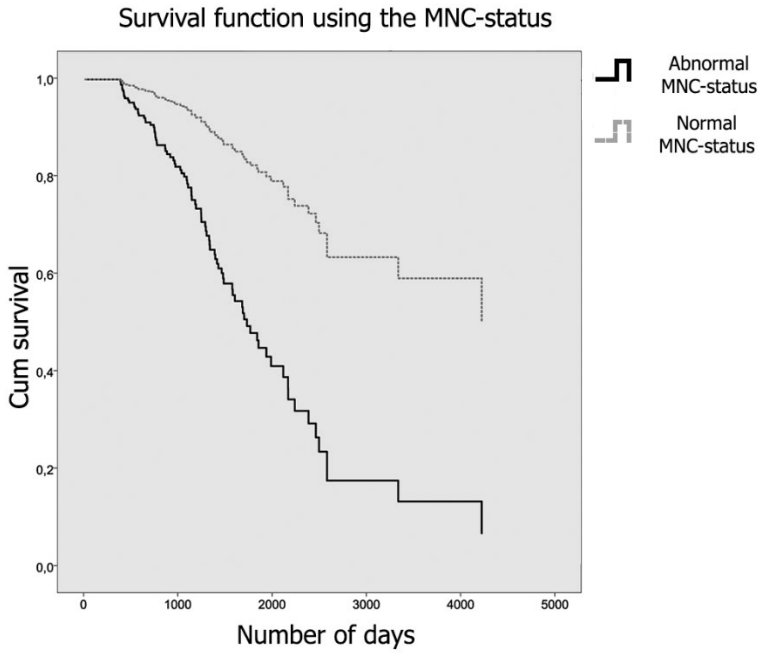


Figure 4. Survival curves plotted for the MNC status. Patients who had an abnormal MNC-status are plotted in a solid line, and patients with a normal MNC-status are plotted in a dashed line. The y-axis shows cumulative survival across time and the number of days is plotted on the x-axis.

4.3.4 Diagnostic overlap

In order to examine whether each method labeled the same individuals as having abnormal cognition, Cohen's kappa's (Cohen, 1960) were calculated, which can be found in supplement 3. There was substantial agreement between the MNC and using 1 abnormal score (cut-off z < -1.45) and 2 abnormal scores (cut-off z < -1.15). There was moderate agreement between using the MNC profile analysis and using 3 abnormal scores with a cut-off score of z < -0.70 (Landis & Koch, 1977).

4.3.5 Added value of the MNC status to MCI status.

Finally, we evaluated whether the MNC profile analysis would add to the predictive value of the clinical evaluation (i.e. MCI vs SCD syndrome diagnosis). To this end, likelihood ratio tests were performed. Block 1 consisted of the syndrome diagnosis (MCI/SCD) and age as predictors of progression to dementia. In the second block we added the MNC status (i.e. normal or abnormal profile based on $p < 0.22$). Adding the MNC status improved the prediction of the conventional MCI diagnosis ($\chi^2(1) = 13.45$, $p<.001$). This means that the MNC status adds unique information about the patient on top of the MCI status and increases the ability to predict dementia at follow-up.

4.4 Discussion

In the current study we evaluated how 'abnormal cognition' should be defined (based on neuropsychological test results) in order to best predict progression to dementia. We did so in a large group of non-demented patients who were diagnosed and followed in an academic memory clinic. We used the Advanced Neuropsychological Diagnostics Infrastructure (ANDI), which allows comparison of each patient to a large representative normative sample (with corrections for age, sex, and level of education) while circumventing problems typically found in neuropsychological practice (outdated norms, no reference data for the old-old, interpretation changes when shifting age groups). During the calibration process we found that when evaluating 10 neuropsychological test measures, the best cut-off for one abnormal score (out of ten test scores) was $z < -1.45$, for two abnormal scores $z < -1.15$, and for three abnormal scores $z < -0.70$. The method of multivariate normative comparison (MNC) predicted decline to dementia best at a cut-off of $p < 0.22$. When testing these cut-off points, we found that they all performed similarly well. Patients who had abnormal cognition by one of these definitions were about four times as likely to progress to dementia compared to the patients who had normal cognition. There was a good agreement between these definitions of abnormal cognition. We also found that even though we used different criteria (one, two, three abnormal test results or abnormal cognitive profile based on the MNC), mostly the same individuals were diagnosed as having abnormal cognition.

Finally, we found that adding the MNC profile analysis to a standard clinical evaluation (i.e. MCI diagnosis) led to a better prediction of which patients would progress to dementia.

The MNC profile analysis predicted similarly well as the univariate criterion methods (one, two or three abnormal scores). Our optimal cut-off score p<0.22 (one-tailed) was far more lenient than when an equivalent of -1.5 SD, i.e. a *p-value of* <0.067 (one-tailed), would be used. This implies that when one looks at a patient's entire profile, certain combinations of small abnormalities may already predict progression to dementia. It also shows that we are far too strict when we use the popular one-tailed *p* value of < 0.05. This would lead to an unnecessarily poor detection of abnormal cognition. Note, however, that when using the MNC profile analysis, it is important to keep in mind which patient population we are assessing. A conventional choice of 0.05 (one-tailed) should probably be used when investigating a population in which cognitive deficits are not a priori expected, for example, when a sample from the general population is screened. But when looking at a group of people with cognitive complaints who present at a clinic because of these complaints, the a priori chances of having some form of cognitive deficit is higher. How much higher is something worth investigating empirically, as we did in our current study. We found that for these memory clinic patients, a rather high *p*-value (0.22 one-tailed) results in the best predictive power. Note that this is a one-tailed p-value, implying that the patients, who are abnormal by this criterion, show a score profile that is worse than that of the 22% of cognitively normal persons with the most negatively deviating profiles.

This study is not the first that aimed at improving the criteria used to define abnormal cognition. Most previous definitions are consensus criteria based on expert opinions. The current study adds to this body of work by providing empirically derived cut-off points. Our results show that the criteria that have been used in the field thus far are justifiable. That is, scores -1.5 SD below the appropriate normative mean, or between -1 and -2 SD are justly considered abnormal (DSM-5, APA,2011, Albert, et al. 2011, Petersen & Morris, 2005). However, our results show that this is the case when one abnormal score out of 10 is used (z < 1.45). When more abnormal scores (two or three) are required, the magnitude of the deviation should be

less strict. Thus, if the number of required deviating scores varies, the magnitude of the deviations should be adjusted accordingly (few deviations require large magnitude of deviations and vice versa).

Bondi et al. (2014) also proposed an improved definition of abnormal cognition that outperformed conventional MCI criteria in predicting which patients would decline to Alzheimer's type dementia. They used six neuropsychological measures and defined abnormal cognition as either a) two measures one SD below the age-corrected mean in one cognitive domain, or b) one such abnormal score in each of three cognitive domains, or c) as functional dependence (but not as much as to qualify for a dementia diagnosis). Adding a functional dependence measure, for example the cognitive functional composite (CFC) questionnaire (Jutten, et al. 2017) creates the benefit of finding individuals who are already compromised in their daily lives and thus will decline to dementia more quickly. We, however, were looking for a way to detect cognitive abnormality at an earlier stage (before functional deterioration sets in). Bondi and colleagues used a cut-off score of -1 SD on two tests in one domain. This criterion is corroborated by the current results, as we found that for two abnormal scores, a cut-off score of $z < 1.15$ gave the best prediction when considering two subnormal scores. For three scores, Bondi and colleagues also used -1 SD as a cut-off score. Our results indicate that in this case a *z*-value < -0.70 may be better.

We investigated whether the MNC profile analysis might be useful as an addition to a standard clinical evaluation. A patient might have an abnormal *profile* without any abnormal test scores. Thus, over and above a conventional MCI diagnosis, information about the patient's cognitive profile is a valuable addition for predicting whether the patient will progress to dementia and will bolster the experience-based insights clinicians already use in their assessments.

This study has some limitations. First, the proportion of individuals labelled as cognitively abnormal depends on the number of tests used. We included 10 test variables. We would like to stress that the optimal cut-off values may be different when a test battery with, say, 5, 15 or 20 test variables is applied.

Second, because only a few tests were used, not all relevant cognitive domains were assessed (for example, language tests were not included). Third, more neuropsychological tests were used in establishing MCI than were used in establishing the MNC-status (as only the tests that overlapped between the patient battery and the ANDI database could be used). For future research it would be worthwhile to set up a study that evaluates the MNC-status with the same tests that were used for the conventional diagnostic procedure. Fourth, the mean age of the patient population was 64 years, which is quite young for a dementia cohort. There were more men (64%) than women, which is also not typical for the dementia population. This is due to the fact that these patients were recruited in an academic hospital specializing in early onset dementia. Moreover, the patients were not demented (at baseline). Therefore, the findings of our study might not be representative for the dementia population as a whole and should be replicated in older samples and in samples with more women. Fifth, we did not investigate different types of dementia. It would be worthwhile for future studies to specifically look at particular dementia types. Finally, we focused only on neuropsychological assessment and did not use neurochemical and neuroimaging biomarkers as predictors. Since combining various types of biomarkers is a powerful strategy of predicting progression to dementia (Virk, Poljak, Braidy, Sachdev, 2018; Olsson et al. 2016; Ruan, D'Onofrio, Sancarlo, Bao, Greco, & Yu, 2016; Blennow, Mattsson, Schöll, Hansson, & Zetterberg, 2015), a future study may extend the current results by incorporating these predictors as well.

In conclusion, when evaluating neuropsychological test results, abnormal cognition can be defined as having either one, two or three abnormal scores as long as the magnitude of the score deviations is adjusted accordingly. This means that when only one abnormal score is required the magnitude of this abnormality needs to be larger compared to two or three abnormal test scores. Using a multivariate profile analysis gives additional information about cognition even when a patient already satisfies the conventional MCI criteria. This profile analysis can, thus, further improve the prediction of progression to dementia.

Chapter 5

Universal Scale of Intelligence Estimates (USIE): representing intelligence estimated from level of education

This chapter has been published as:
N.R. de Vent, J.A. Agelink van Rentergem, M.C. Kerkmeer, H.M. Huizenga, B.A. Schmand,& J.M.J. Murre (2016). Universal Scale of Intelligence Estimates (USIE): representing intelligence estimated from level of education. *Assessment, 25*(5), 557-563

Abstract

In clinical neuropsychology it is often necessary to estimate a patient's premorbid level of cognitive functioning in order to evaluate whether his scores on cognitive tests should be considered abnormal. In practice, test results from before the onset of brain pathology are rarely available, and the patient's level of education is used instead as an estimate of his premorbid level. Unfortunately, level of education may be expressed on many different scales of education, which are difficult to use interchangeably. Here, we introduce a new scale that has the capacity to replace existing scales and can be used interchangeably with any of them: the Universal Scale of Intelligence Estimates (USIE). To achieve this, we will map all levels of existing educational scales to standard IQ scores. A USIE point estimate is supplemented with an estimation interval. We assert that USIE offers some important benefits for clinical practice and research.

In most clinical neuropsychological assessments, some estimate of the level of patients' cognitive functioning prior to the onset of their cognitive complaints is required to decide whether a certain test score is abnormal (Lezak, Howieson, Bigler, & Tranel, 2012). Ideally, this level would be summarized by a premorbid IQ score. In practice, however, such a score is rarely available. Instead, the patient's level of

education is used as a proxy for his or her (premorbid) level of cognitive functioning. In neuropsychological research, it is also common to use level of education as an indicator of cognitive functioning, for example, as a covariate in statistical analyses.

This paper discusses problems that limit the usefulness of level of education as a proxy for premorbid cognitive functioning. We propose a remedy for these problems in the form of a new scale: the Universal Scale of Intelligence Estimates (USIE). Below, we discuss how level of education is currently used in the clinical setting to estimate premorbid cognitive functioning and the limitations of this approach. Then we will introduce the concept of USIE, its construction, and its benefits. Finally, we will discuss some practical issues related to the construction of USIE and some possibilities for a broader application. First, however, we will briefly review some other solutions that have been proposed for estimating premorbid IQ.

5.1 Estimators of premorbid cognitive functioning in clinical practice

In the absence of actual data on premorbid cognitive functioning, an estimate may be derived from cognitive tests that are relatively insensitive to brain pathology. In this section we will review a number of methods that are used in clinical practice to acquire an estimate of premorbid cognitive functioning.

5.1.1 Cognitive tests for premorbid IQ estimations

Tests such as the National Adult Reading test (NART) (Nelson, & O'Connell, 1978) and its US adaptations the North American Adult Reading Test (NAART) (Blair, & Spreen, 1989) and the American version of the National Adult Reading Test AMNART (Grober, Sliwinsk, Korey, 1991), are tests that rely on word-reading abilities which are relatively spared in many forms of brain pathology and which are used to estimate a patient's premorbid IQ. Similar tests are the Wide Range Achievement Test-Word Reading (WRAT-READ) (Wilkinson & Robertson, 2006) or the Wechsler Test of Adult Reading (WTAR) (Wechsler, 2001) which has norms corrected for level of education. A variation on these word reading tests are tests that use the force-choice format such as the Spot-the-Word Test (STWT) (Baddeley, Emslie & Nimmo-Smith,

1993) or the Lexical Orthographic Familiarity Test (LOFT) (Leritz, McGlinchey, Lundgren, Grande, & Milberg, 2008).

Another way to estimate premorbid cognitive functioning is by using information about a patient's background in addition to his level of education. For example, using a regression-based approach to estimate premorbid IQ from demographic variables such as age, sex, race, level of education, occupation and region of the country (Barona, Reynolds, Chastain, 1984; Crawford, & Allan, 1997). A combination of using neuropsychological test measurements and demographic variables is used by the Oklahoma Premorbid Intelligence Estimation (OPIE-3), which takes into account both the demographic background of a patient as well as WAIS-III test data from the subtests Vocabulary, Information, Matrix Reasoning and Picture Completion (Schoenberg, Duff, Scott, Patton, & Adams, 2006).

There are a number of practical limitations to these methods. First, the word-reading tests cannot be used in patients with more severe pathology or with a language deficit (such as aphasia or dyslexia). A similar problem occurs with the OPIE-3 estimate as not all patients will be able to reliably complete these WAIS-III subtests. Second, the regression-based approaches may be satisfactory in the USA or UK where they were developed, but they cannot easily be applied in other countries as the formulas are based on the demographic characteristics of the USA (or UK). Lastly, in large scale research projects it is not always possible to test each participant face-to-face and therefore gathering a measurement of premorbid IQ is not always a possibility.

5.1.2 Level of education as an indicator for premorbid IQ

When measurement of premorbid IQ is not possible, level of education is often used as an estimate of a patient's premorbid level. This is universally applicable, also in case of severe impairments. Moreover, the normative data of many neuropsychological tests are stratified by level of education whereas only few tests can be corrected for (premorbid) IQ (Lezak, et al., 2012). This has led to frequent use of 'level of education' to represent a patient's premorbid cognitive abilities in the diagnostic process.

5.1.3 Limitations of level of education

Even though level of education is widely used as an indicator of cognitive functioning, it entails a number of problems. Some are of a conceptual nature, whereas others are more practical.

Many educational scales: Even in one country there may be several ways of defining level of education. It may be quantified as number of years of formal education, or in terms of the highest intellectual level of completed education (e.g., 'college' or 'elementary school'). Comparing values measured on different scales cannot easily be done and comparing measurements between countries is even more difficult as a particular school type may be unknown in another country. The same scale point, such as 'college', may represent a higher intellectual level in one country than in another. In an attempt to remedy this, UNESCO introduced the International Standard Classification of Education (ISCED; see table 1) (UNESCO, 1997), which codes level of education on a 7-point scale. A more recent ISCED version has a more differentiated scale (UNESCO, 2011). In contrast to this rather fine-grained scale, a rough division into three groups, typically ‘low’, ‘medium’, and ‘high’, is also widely used. Some researchers even decide to construct their own scales for educational levels.

The many different educational scales may cause great difficulties when researchers want to compare data from different studies, for example in a meta-analysis, if level of education is a variable of interest. Different coding systems may prove impossible to translate into a single, all-encompassing system. For example, when one study uses the ISCED (1997) scale and another study uses education in years, the studies are difficult to compare because there is no straightforward conversion from ISCED (1997) to years of education or vice versa. This problem is often resolved by recoding the finer-grained scales into a scale with a lower resolution with categories such as a ‘low’ ‘medium’, and ‘high’. Although such recoding is feasible and easy, it causes loss of information about participants’ level of education, which in many situations is undesirable.

Table 1: *Level of completed education and corresponding score on the ISCED (1997) scale.*

ISCED score	Level of completed education
0	Pre-primary education
1	Primary Education
2	Lower secondary
3	(Upper) secondary
4	Post-secondary, non-tertiary
5	First stage of tertiary education
6	Second stage of tertiary education

Outdated educational scales: Some educational scales have become outdated due to changes in educational systems in the course of time. For example, when levels of education or entire school systems are redefined or reformed, the new school system may not fit into the existing scales. To solve this problem, new scales must be created, or the new levels of education have to be forced into scales that were not designed to accommodate them. This poses a challenge for clinical use when test scores are compared to a norm table that is stratified by level of education, because it may introduce inaccuracies. Ideally, norm tables should contain all possible levels of education that are in use today, but also those that were in use when elderly patients went to school, so that we can compare our patients to the correct normative sample. In practice, test norms never include such extended educational tables.

Level of education as an ordinal scale: In research settings, level of education is often used in statistical analyses. Educational levels are coded on an ordinal scale, such as ISCED, and numerically recoded, after which statistics such as average and standard deviation are calculated. On an ordinal scale, however, statistics such as average, standard-deviation and correlation are not well-defined, and using them in this manner may give a distorted impression of the population.

5.2 USIE – Universal Scale of Intelligence Estimates

To alleviate these problems, we propose that existing scales of education are replaced by a single universal scale, which uses estimated IQ scores instead of educational level: the Universal Scale of Intelligence Estimates (USIE). The USIE presents mappings from existing educational scales to a (pseudo) IQ score. For example, the mapping of the ISCED (1997) to the USIE can be achieved by collecting IQ scores for each of the seven points on the ISCED scale. If this is done for a number of different educational scales, we can express each scale in terms of USIE IQ scores. In this manner, all educational scales can be replaced with a single USIE IQ scale, which is used as an intermediate scale. For example, an educational scale with scale points 'low', 'medium', and 'high' might map to corresponding USIE scores of 91, 97, and 111, respectively. An ISCED (1997) score of 5 might map to 116 (see Figure 1). There are some important advantages associated with the USIE IQ scale, provided the necessary data for its construction are available. The next section will focus on the data required for the construction of the USIE.

5.2.1 Data required for USIE construction

In order to construct USIE, data are required from people whose level of education and IQ have been measured. For example, if we wish to map the ISCED (1997) scale to USIE, we would gather IQ data for each of the 7 points on the ISCED scale. Using the IQ format implies that by definition an IQ score of 100 is equal to a USIE score of 100.

A straightforward approach is to treat level of education as a nominal variable; the USIE scores are tabulated as the median IQ scores for the different levels of education. Medians are used instead of means because IQ scores may not be normally distributed at every educational level. We propose to standardize USIE on the Wechsler Adult Intelligence Scale Fourth Edition (WAIS-IV) (Wechsler, 2008), because this is the latest edition of the most widely used IQ scale. Moreover, editions for many languages and countries are available. To illustrate USIE's construction procedure we used data from the Dutch edition of the WAIS-IV (Wechsler, 2008).

5.2.2 Precision of the estimate

An estimation interval surrounding the USIE score can be constructed corresponding to, for example, the 10th and the 90th percentile of the IQ distribution obtained from the population. The width of this interval is influenced by the resolution of the educational scale being mapped. When a scale has a high resolution, USIE values will tend to have relatively narrow estimation intervals. This means that there is relatively little variation in IQ between participants who have the same score on this educational scale. When an educational scale is quite crude (say a 3-point scale), there will generally be more variation in IQ scores between participants who have the same score on this educational scale. This leads to a somewhat higher between-participant variance, and thus to larger USIE estimation intervals.

5.3 USIE application

5.3.1 Comparing educational scales through USIE

When two different educational scales have been mapped to USIE, we can compare both scales and we can visualize their relationship. For this example we have used data from the Dutch WAIS-IV. This data consisted of 1172 participants of which 50.9% were female. The data was gathered from August 2010 to May 2011. More detailed information on this dataset can be found in the Dutch manual of the WAIS-IV (Wechsler, 2008). Under the assumption that the data from the WAIS-IV has been randomly sampled, we can now draw a figure such as Figure 1. In Figure 1, the first educational scale is the 7-point ISCED (1997) scale, the second scale has only three categories, namely ‘low’, ‘medium’, and ‘high’. From this figure it becomes apparent that someone who has a ‘medium’ level of education has a USIE score of 97, which is roughly the same as an ISCED (1997) score of 3, which corresponds to a USIE score of 100. USIE can thus also function as an intermediary scale. Surrounding each USIE point estimate is an 80% estimation interval. This information can also be found in table 2.

Table 2. *USIE estimates per level of education on the ISCED and low/med/high scales*

Level of education		USIE	10th percentile	90th percentile	N
ISCED	0	-	-	-	0
	1	91	53	118	171
	2	92	75	108	322
	3	100	83	118	393
	4	108	94	124	190
	5	116	101	133	89
	6	-	-	-	0
Low/Med/High	Low	91	53	118	162
	Medium	97	79	114	711
	High	111	95	127	277

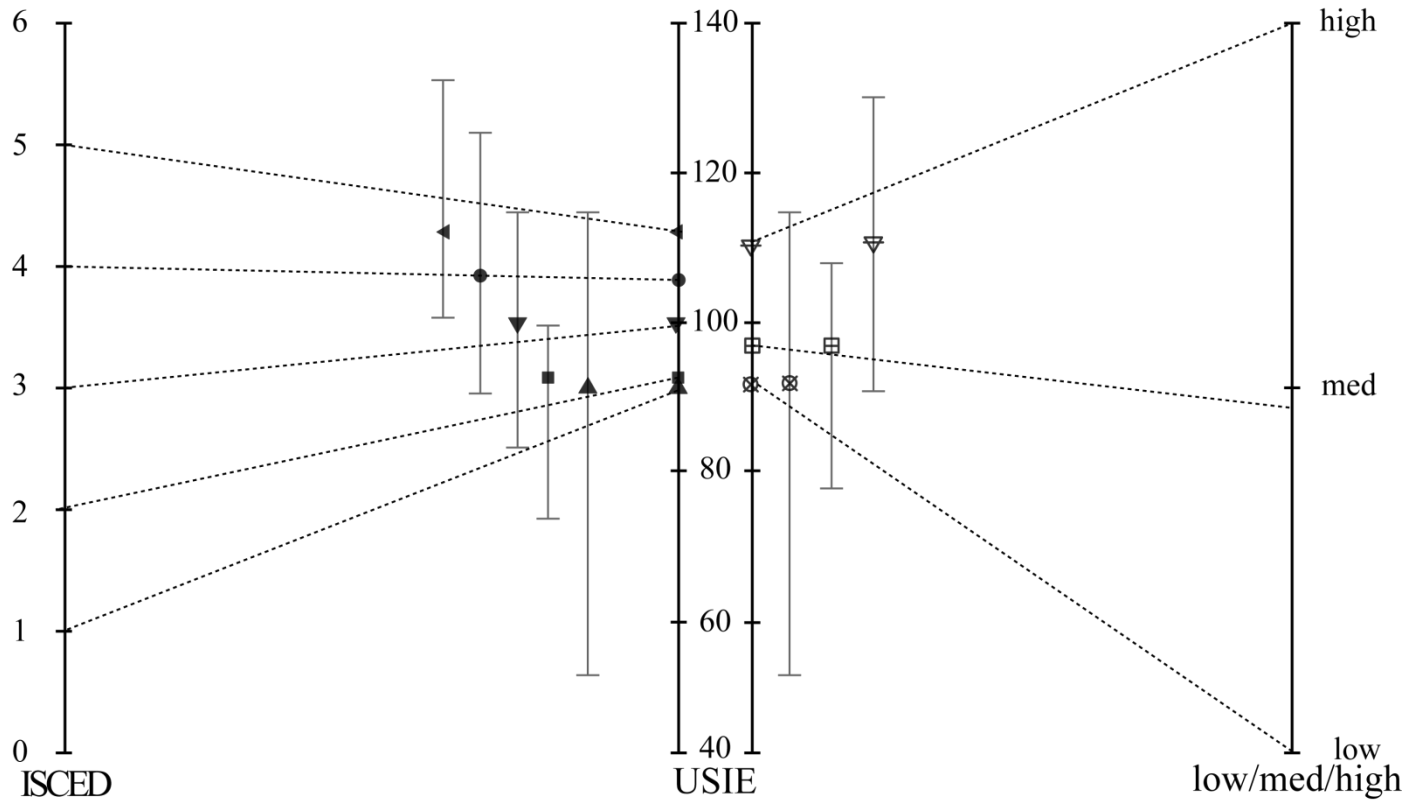


Figure 1. Mapping of the ISCED (1997) scale and an educational scale with lower resolution ('low', 'medium', and 'high' education) to the USIE scale. For each educational level the median WAIS IV IQ and its corresponding 80% estimation interval (i.e. 10^{th} and 90^{th} percentile) are plotted as USIE values. Note that the '0' and '6' scale points of the ISCED (1997) do not have a corresponding USIE; there are no observations in these categories.

5.4 Discussion

5.4.1 Advantages of USIE

In this paper we introduced the Universal Scale of Intelligence Estimates (USIE) scale as a new way to represent cognitive level as estimated from level of education. This new scale does not solve all problems with current approaches to estimating (premorbid) cognitive level or intelligence, but we have argued that it does alleviate some of them. In this section, we will briefly summarize the advantages. In the next section, we will discuss limitations that remain and ways to possibly overcome these.

Many educational scales: By far, the largest advantage of the USIE scale is that it allows the translation and comparison of scores on different educational scales. If two scales, such as ISCED (1997) and years of education, are mapped to USIE, meaningful comparisons become possible. For example, a person with an ISCED (1997) score of 2 may have a USIE score of 90, and a person with 12 years of education may have a USIE score of 105. This was illustrated in figure 1.

Outdated educational scales: Because of changes in the educational system, some levels of education that were common decades ago no longer exist today. With USIE scores this problem can be circumvented. To achieve this, it is necessary to construct mappings of both the outdated and current educational scale(s) to USIE. This also implies the advantage of not having to update test manuals when new educational systems are formed: one can simply measure the IQ of those who have completed the new type (and level) of education.

Level of education as ordinal scale: The USIE is conceived as an IQ score, which is represented on an interval scale. All existing educational scales known to us are ordinal. Converting educational scales into the USIE interval scale allows valid calculation of statistics such as the mean and standard deviation, which are not defined on an ordinal scale.

USIE as predictor of premorbid IQ: Even though the primary application of USIE is to map and compare different scales of education, it can also be used as an estimate of premorbid IQ. Unlike word-reading tests such as the NART or the WRAT, it is completely insensitive to brain pathology. Therefore, it can be used even in the most severely affected patients. When comparing USIE to regression-based estimations of premorbid IQ (such as the Barona method or the OPIE-3), USIE is specifically useful when not all required demographic variables are available to use the other methods. This is particularly the cases when comparing participants between studies such as in a meta-analysis in which not all studies collected the same demographic variables. The USIE only requires information about a participant's level of education, something that is often available. Also, these regression-based approaches have to be calibrated for each specific country because the coefficients that are estimated in one country are not automatically applicable in another country. Also they would have to be regularly recalibrated because the demographic characteristics of a country usually change over time. The USIE will also have to be recalibrated, but this can more easily be done when publishers of IQ tests incorporate information about the level of education of their normative sample. Lastly, the USIE is accompanied by an estimation interval which gives the user extra information about the certainty of the USIE estimate.

5.4.2 Limitations of USIE

As mentioned before, there are many different scales of education. Creating a mapping to USIE for each of those scales is costly because it is labour-intensive and time-consuming to collect sufficient IQ data for all educational scales. However, this investment will be worthwhile, because it needs to be done only once for each scale of interest. Also, if the exact level of education and educational degree are collected for

each participant, it would allow mapping to USIE from every imaginable educational scale. For example, if we know that participants have completed a master of science degree, we can use their IQ data to map the ISCED level 5 (first stage of tertiary education), but also to map the 'high' category of the 3-point 'low', 'medium', 'high' educational scale. We would have to collect such fine grained data only once to make a wide range of mappings to USIE available.

Based on the same principles for comparing different educational scales within one country, USIE could theoretically also be used to compare educational levels between countries. However, we need to be cautious when doing so, because USIE scores are expressed as IQ scores. Consequently, they are dependent on the population characteristics of the country or (sub) culture in which the IQ test was standardized. However, USIE can probably be applied to compare educational levels in countries that are sufficiently similar, such as most Western countries. These applications are non-trivial, and we expect the USIE scale will offer a significant improvement, for example, when conducting a study in various European countries, or when conducting for example, a meta-analysis that includes studies from Europe and the United States.

The USIE scale is not influenced by brain pathology. It is based on a demographic factor instead of measurements taken after acquiring brain pathology. This implies, for example, that for a patient with suspected Alzheimer's disease, the USIE scale can give a fair indication of his premorbid cognition because the patient completed his school education prior to the onset of the disease. On the other hand, this also points to a limitation of USIE, namely its reliance on a fairly normal school career. Because the USIE scale is based on the level of education, premorbid IQ will be underestimated for patients who have had a suboptimal school career, for example due to disease or other personal circumstances. This makes USIE unsuitable for application with developmental or congenital disorders such as ADHD, autism, or schizophrenia (see Dennis, Francis, Cirino, Schachar, Barnes, & Fletcher, 2009).

5.4.3 Possibilities of improving or extending the USIE concept

First, for a scale like USIE to be really useful, norm tables for clinical tests need to be specified by (USIE) IQ scores instead of by level of education as is currently the case. Even better would be to replace existing norm tables with regression based norms, where USIE can be included as a covariate (Crawford & Allen, 1997; Testa, Winicki, Pearlson, Gordon, & Schretlen, 2009).

Second, when level of education is used as a proxy for premorbid cognitive abilities, one has to be aware that education by itself only has limited correlation with cognitive functions. For example, education and IQ correlate in the order of 0.55 to 0.65 (e.g., Barona, Reynolds, & Chastain, 1984; Crawford & Allan, 1997; Crawford, Millar, & Milne, 2001; Harnett, Godfrey & Knight, 2004; Matarazzo, 1972; Wilson, Rosenbaum, Brown, Rourke & Grisell, 1978). In research and clinical practice, however, level of education is often accepted as a-good-enough estimate of general cognitive functioning. This could lead to underestimation of a patient's premorbid level of functioning. For example, premorbid level might be underestimated in case of elderly women who have had few years of formal education because they were not given the opportunity to pursue a higher level of education. Their level of education may be quite low while their IQ's are (above) average. Researchers and clinicians can keep this in mind, and possibly create separate USIE estimates for men and women in specific age bins.

5.4.4 Concluding remark

Level of education has long been used to estimate (premorbid) level of cognitive functioning, but the variation in educational scales hampers research and clinical practice. We have argued that the USIE scale will facilitate meta-analyses and other types of research and will better support clinical decision making. Clearly, much work needs to done to map educational scales to USIE. Once this has been completed, meta-analyses will no longer be hampered by diverging scales of education, and comparison of cognitive studies in different Western countries will become more feasible.

Chapter 6

Summary and Conclusions

6.1 Summary

The goal of this thesis was to help alleviate problems typically found in neuropsychological assessment and to even further improve on this process. We mainly focused on improving normative comparisons which are performed to evaluate patients' test scores and to decide whether a score is either normal or abnormal. Firstly, we created a large normative database that allowed corrections for age, sex, and level of education, so that each patient could be compared to the relevant set of healthy control participants. Secondly, we applied a multivariate normative comparison method which enables a formal profile analysis on all demographically corrected test scores of patients.

Chapter two described the construction of the aggregated database. By combining generously donated healthy participant datasets, we were able to construct a normative database without testing participants ourselves. Data from healthy participants came from control subjects in clinical studies, as well as from large community-based studies. In order to make a large, coherent database from multiple studies, a standardized procedure was followed. (1) Extreme borders: In order to make ANDI sensitive for detecting abnormal performance, we needed to remove scores that could potentially endanger this detection ability. We did this in two steps. First, we checked whether scores obtained from participants were within the limits of the test. If scores were impossibly high (e.g. higher than the perfect scores) these were removed. With the help of the clinical expertise of our ANDI steering committee we also defined a lower bound for each test. If participants scored below this defined score, they could no longer be assumed cognitively healthy, and their scores were also removed.

(2) Model selection: We decided which demographic corrections were necessary by using the Akaike Information Criterion (AIC). (3) Removing outliers: Based on the output from the AIC we removed demographically corrected outliers, i.e. scores that were very low given a patient's sex, age, and level of education. (4) Normality: In order to be able to use parametric tests and because many of the neuropsychological test data were not normally distributed, we transformed these scores to normality. For each test variable, the Box-Cox procedure (Box & Cox, 1964) was used to find the best possible power transformation, which was then used to transform the data. (5) Refit models: With the cleaned and normalized database we decided which regression weights for the demographic variables we had to implement on the ANDI web portal. For this we used the AIC again. The last section of chapter two described a selection of the contents of the ANDI database. For each variable we described the number of donated datasets, the number of healthy individuals, age range, % men, and education range. For example, the Auditory Verbal Learning Test (AVLT) delayed recall variable had been donated by 29 studies, had 4598 healthy individuals, age ranged from 14-97, 49% men and education ranged from 1 (unfinished education) to 7 (university degree).

In chapter three, the ANDI database and the multivariate normative comparisons method were put to the test by re-analyzing longitudinal data from a study in patients with newly diagnosed Parkinson's disease (PD), part of whom developed Parkinson's disease dementia (PDD) (Broeders et al., 2013). Previously, conventional criteria for Parkinson's disease Mild Cognitive Impairment (PD-MCI; Litvan et al., 2012) were used to predict progression to PDD (Muslimovic et al., 2005; Muslimovic et al, 2009; Broeders et al, 2013). In the current chapter, we investigated whether using ANDI and the multivariate normative comparisons would improve the prediction of progression to PDD. First, we looked at the PD-MCI criteria in the conventional (univariate) way, but with the normative scores of the ANDI database. That is, we applied the PD-MCI criteria but instead of using traditional normative data we used the ANDI database and its demographically corrected norms. This showed that fewer patients were labelled as having abnormal cognition by the ANDI database compared with the traditional norms and criteria.

This was true for both the patients who progressed to PDD as well as the patients who remained cognitively stable. The sensitivity and specificity were similar for the PD-MCI criteria with traditional norms and with ANDI norms (e.g. the ANDI database did not improve the prediction). Secondly, we looked at whether using the multivariate normative comparisons with the ANDI database would result in a better prediction of which patients would progress to dementia compared to the conventional PD-MCI criteria. This appeared to be the case. Both sensitivity and specificity were higher for the multivariate normative comparison method than for the conventional PD-MCI criteria (with traditional norms and the ANDI database). The results from this study indicated that both the multivariate normative comparisons and the ANDI database allowed for a more sensitive prediction of PDD than the traditional PD-MCI criteria.

In chapter four, ANDI and the multivariate comparison method were used to help define 'abnormal cognition' in a group of non-demented patients from the Amsterdam Dementia Cohort (van der Flier & Scheltens, 2018). Several definitions of abnormal cognition have been given in the literature but there is still no consensus. In this chapter we investigated how to best define abnormal cognition when predicting progression to dementia, and whether the multivariate normative comparison method could improve this prediction. First, using a cross validation method, we determined which number of abnormal tests and which magnitude of score deviations best predicted progression to dementia. Using 10 neuropsychological test measures, we found that the predictive ability for progression to dementia of one, two and three abnormal test scores is highly similar, provided that cut-off values are adapted appropriately. That is, when one deviating score is required the deviation has to be larger than when three deviating scores are required. We also found that a multivariate profile analysis gives additional information about a person's cognition even when he/she satisfies the conventional MCI criteria. This profile analysis can thus further improve the prediction of progression to dementia.

Chapter five took a side-step and looked at the proxies for premorbid intelligence that are used in neuropsychological evaluation. Because true premorbid IQ scores are almost never available for patients, we generally make use of a proxy, for example tests that are less vulnerable for neuropsychological problems. Also, an estimate of premorbid intelligence may be obtained by asking about a patient's background such as

job history or level of education. Level of education is the most often used proxy, as it is almost always available. However, it is expressed on many different scales, ranging from high resolution to very crude (low, medium, high), and this makes comparing them (for example in a meta-analysis setting) difficult. We propose a new scale for premorbid IQ that can be used interchangeably with all others: The Universal Scale of Intelligence Estimates (USIE). The USIE presents mappings from different existing educational scales to (pseudo) IQ scores. When this is done for a number of different educational scales, we can express each scale in terms of USIE IQ scores. In this manner, all educational scales can be replaced with a single USIE IQ scale, which is used as an intermediate scale. If all educational scales are mapped to the USIE, the USIE will facilitate meta-analyses and other types of research and will better support clinical decision making.

6.2 ANDI's contribution to the diagnostic process

ANDI offers a number of important improvements for normative comparisons in neuropsychological practice. First, the benefits of data sharing have been illustrated by ANDI. In our current climate in which many sciences are economized it is important to use what was already gathered in order to optimize our chances of realizing progression in our fields of study. By making use of a large aggregated database ANDI made it possible to have a representative control sample for each patient.

Second, ANDI introduces a regression-based normative database instead of a more traditional tabular approach. This makes it possible for clinicians to better evaluate a patient's cognition. The use of regression analysis for creating norms gives a number of advantages over using traditional norm tables which we will discuss here. The norm tables use one or more demographic variables (mostly age, sometimes age and education, or age and sex) and give normative statistics such as means, standard deviations, percentiles or standard scores for each subgroup. The borders, e.g. on age, of each subgroup are determined by the publishers and thus can widely vary between tests. For example, on test A patient of age 63 could be in the 50-70 age group while on test B he might be in the 60-80 age group. Compared to the other individuals in the group, the patient would perhaps score better on test B than on test A purely as a result of how the norm groups are defined. Also, when a patient becomes older and moves to another subgroup, the

interpretation of his test scores may differ. Also, this traditional approach using norm tables requires many data, because each of the defined groups needs to be represented by a sufficiently large sample. A regression-based approach requires about one-third of the sample size that is required for traditional norms (Oosterhuis, van der Ark, & Sijtsma, 2016). This approach of regression-based norms was already successfully introduced in 1985 (Zachary & Gorsuch, 1985) and applied in neuropsychology by Crawford and Howell in 1998. They have also worked on several methods to best evaluate the difference between the patients' predicted score (obtained with the regression equations) and their observed score. Using their methods to define for example, more correct confidence intervals in smaller normative samples might be something ANDI can incorporate in the future, if tests with small normative samples become available in ANDI.

Third, ANDI is analyzing test results in a multivariate or co-normed fashion. This trend has increased in recent years. For example, the latest versions of the Wechsler test batteries have also been co-normed (WAIS-IV, WMS-IV). This makes it possible to compare a patient's test score to other scores he has obtained (as the correlation between tests is known). This way something can be said about a patient's profile score. But the application is limited to the Wechsler test materials. Other initiatives such as the MATRICS Consensus Cognitive Battery (MCCB) for cognitive impairment in schizophrenia (Nuechterlein et al, 2008), EXAMINER which measures executive abilities (Kramer et al, 2014), and the NIH toolbox for assessing neurological and behavioral function (Gershon et al, 2010) have all brought progress to their respective fields by creating large databases for neuropsychological tests and evaluating tests in the context of other tests. ANDI on the other hand introduced the statistical method of multivariate normative comparison (MNC). This method has shown to improve the sensitivity and specificity for detecting abnormal cognition (Huizenga et al., 2007) In this thesis we have shown that it performs better than traditional univariate methods for a group of Parkinson's patients as well as patients with memory complaints. The MNC statistically sharpens the intuition clinicians already have about the cognitive profile of their patients. For example, when a patient completes a memory test with two conditions, an immediate recall and a delayed recall, we traditionally evaluate both conditions separately. We compare the scores on each condition to the corresponding norm tables and evaluate whether the scores are normal or abnormal. But a clinician's intuition goes further than

looking at these test results. Clinicians know that when a patient has a certain immediate recall score, they are expected to have a certain corresponding delayed recall score. So, when a patient's immediate recall is very good and thus shows a high learning capability, the delayed recall is expected to be good as well. When the norm tables indicate the delayed recall, score is on the low end of the distribution but not yet abnormal, clinicians will not easily dismiss this patient with the conclusion he has no cognitive problems (even if no statistically significant abnormality is found). Clinicians look at patients in a clinical context. So, a high score on immediate recall should be accompanied by a similarly high score on delayed recall. The multivariate normative comparison method formalizes this clinical intuition as it evaluates the highs and lows in a patient's test profile.

Fourth, The ANDI procedures are available for clinicians and researchers to use in a free and user-friendly tool (www.andi.nl). Users can fill out the test scores of their patients and the ANDI web portal will give a detailed report on the patients results. This makes using the advantages of ANDI easy and accessible for all neuropsychologists.

Fifth, ANDI can be used to create control groups for clinical studies. Researchers do not have to gather al the control subjects for their clinical studies themselves. They can simply analyze their clinical data with the healthy reference sample supplied by ANDI.

Sixth, all information about the construction of the ANDI database and the normative comparisons procedures is freely accessible (www.andi.nl) By doing so we hope to promote replicability of this process, whether it concerns creating a database of neuropsychological test scores in other countries, other fields of psychology, medical science and specialties, or still other fields.

6.3 ANDI's shortcomings and possible improvements

Despite ANDI offering important improvements to the field of neuropsychology and normative comparisons in general, there are some shortcomings, which will be addressed here. First, because the data

for the database were donated by a diverse set of universities and (academic) hospitals throughout the Netherlands and Flemish Belgium, variations in the data might have occurred. Even though all data were gathered by trained personnel, some variations in test instructions may have resulted in variance in test scores not accounted for by the demographic corrections. This can cause a less optimal norm sample and thus might have influenced the sensitivity and/or specificity of the ANDI database. We accounted for this variation between studies by means of a multilevel analysis. After extensively testing the ANDI database with (fictional) patients we have not found any indication that this variation between donations is a problem. This variation could also be beneficial as clinicians in daily life do not always stick to same protocols. The data in the ANDI database are thus a good reflection of the variation in neuropsychological test administration

Second, to make the data processing for the ANDI database uniform and easy to understand we chose to only use linear regression models and only include main effects. We do however, recognize there can be some problems with this approach. In some instances, the assumptions for linear regressions might be violated, which could negatively influence the quality of the norms. For example, it is widely known that the effect of age on neuropsychological tests usually does not follow a linear curve (Lezak et al., 2012; Verhaegen & Salthouse, 1997; Salthouse, 2000). We also only used the main effects of age, sex, and level of education as predictors of neuropsychological test scores. Possible interaction effects were not used in the current models. If the ANDI database further increases in size, we wish to revisit these methods to see whether using non-linear regression approaches and the inclusion of interaction terms may lead to a better normative comparison.

Third, when exactly an MNC profile score should be deemed abnormal remains under investigation. In chapter three we investigated how the conventional 5% criterion (5% or less of the healthy population has obtained this score) predicted progression to Parkinson's disease dementia. We showed that the 5% MNC procedure predicted progression to dementia with higher sensitivity and specificity than the traditional consensus criteria-based method. In chapter four, we calibrated a specific cut-off score for a group of memory clinic patients. This cut-off score appeared to be far more lenient than 5%. It was 22%, i.e. a score obtained by 22% or less of the healthy population. Both the Parkinson's and the memory clinic groups were

individuals with no, or only mild, cognitive impairments, and their mental status was quite comparable (the Parkinson's disease group and the memory clinic group both had a mean MMSE of 28). This could be taken to suggest that when a more lenient critical value would have been used in the Parkinson's disease group, the neuropsychological test scores would have predicted decline to dementia even better. It will thus be worthwhile to further investigate which cut-offs for the MNC will work best for which clinical sample. In a group of patients with severe cognitive difficulties, an even more lenient criterion perhaps works better to correctly identify all patients who, for example, will progress to dementia.

Fourth, another challenge of the multivariate normative comparison method is that it requires normative data on co-normed tests (i.e. the correlations between the different neuropsychological tests need to be known). For the current ANDI database, we have co-normed data available for all tests. But if we wish to include more tests, we need to know the correlations between the new (to be added) and the old tests as well. Obtaining these data can be a challenge. Other options worth investigating are using a factor model to estimate missing correlations in the multivariate data. If we consider three memory tests A, B and C and the correlations between A and B, and B and C are known, then the correlation between A and B can be estimated by using the latent (memory) factor. (Agelink van Rentergem, de Vent, Schmand, Murre, & Huizenga, 2017).

Fifth, in the patient studies reported in this thesis, we used the ANDI database and methods to reanalyze existing data sets. This is not an optimal approach as the tests that were used in these datasets did not always overlap fully with the ANDI database. For example, the neuropsychological evaluation for the memory clinic patients was more extensive than the 10 variables reported in this thesis, but the other tests that were used are not in the ANDI database and thus could not be analyzed. Vice versa, the ANDI database had more tests available that were not administered to these patients. For future research it would be very valuable to conduct studies in which ANDI is incorporated in the design stage of the study. This way a more optimal overlap between the neuropsychological tests can be acquired and ANDI can be used and investigated to its full potential.

6.4 Wider application of the ANDI concept

A longitudinal solution: The current ANDI database is mostly used by clinical neuropsychologists who want to determine whether scores their patients obtained are abnormal given their demographic background. However, many clinicians follow their patients over time. Ideally therefore, we would also build a database with longitudinal data to enable clinicians to evaluate whether the *change* over time is abnormal or not. Data are already available for some tests and for certain follow-up moments such as six months, one year or two years. These data are however only applicable if clinicians also test their patients at the same interval, which is not always the case. A better option for creating longitudinal norms would be gathering follow up data over different time intervals and creating regression slopes for performance over time for each of the tests in the ANDI database.

Clinical profiles: At this moment ANDI only has data from cognitively healthy participants as it was designed to be able to detect cognitive abnormalities. In the future, data on specific cognitive disorders could also be used to create pathology profiles. This way the ANDI normative database can first evaluate whether a patient has an abnormal pattern of scores, and then compare this abnormal pattern to a database of various cognitive pathologies. This way ANDI can give clinicians a suggestion what possible brain pathology a patient might have.

Evaluate prior knowledge about patients: Currently, clinicians have intuitions about their patients based on the information they have about the patient's cognition and history. The scores obtained from the patient are not interpreted as stand-alone results but are evaluated with the prior history of a patient in mind. These 'priors' that clinicians have could perhaps also be formally evaluated using a Bayesian approach.

Expanding to other fields of psychology: What we did with the ANDI database for clinical neuropsychology can theoretically be done for any field of study that compares characteristics of a single case to some reference data. In the future, we would like to expand ANDI to other fields of psychology such as healthcare/clinical psychology and developmental psychology, if the appropriate data can be obtained. In many of these fields tests overlap. For example, when psychiatrists have to plan a treatment for their

patients with schizophrenia, they might want to evaluate cognition. Or when a patient comes in with severe behavioral problems, they might need to assess whether this is a psychiatric disorder or perhaps (fronto-temporal) dementia. Having access to norms that span several fields of study will be a valuable contribution to psychology and psychiatry, and indeed to medicine as a whole.

6.5 Conclusion

It is of utmost importance that neuropsychologists are able to compare their patients to a representative healthy norm sample. This entails comparing in terms of age, sex, and level of education. This promotes more reliable detection of those patients who are cognitively abnormal and who require some form of intervention or care. The ANDI database has shown to improve normative comparisons by creating more representative norms and by offering an improved statistical framework.

Chapter 7

Supplemental Materials

7.1 Supplemental Materials Accompanying Chapter 3

7.1.1 Chapter 3 supplement 1: Score profiles

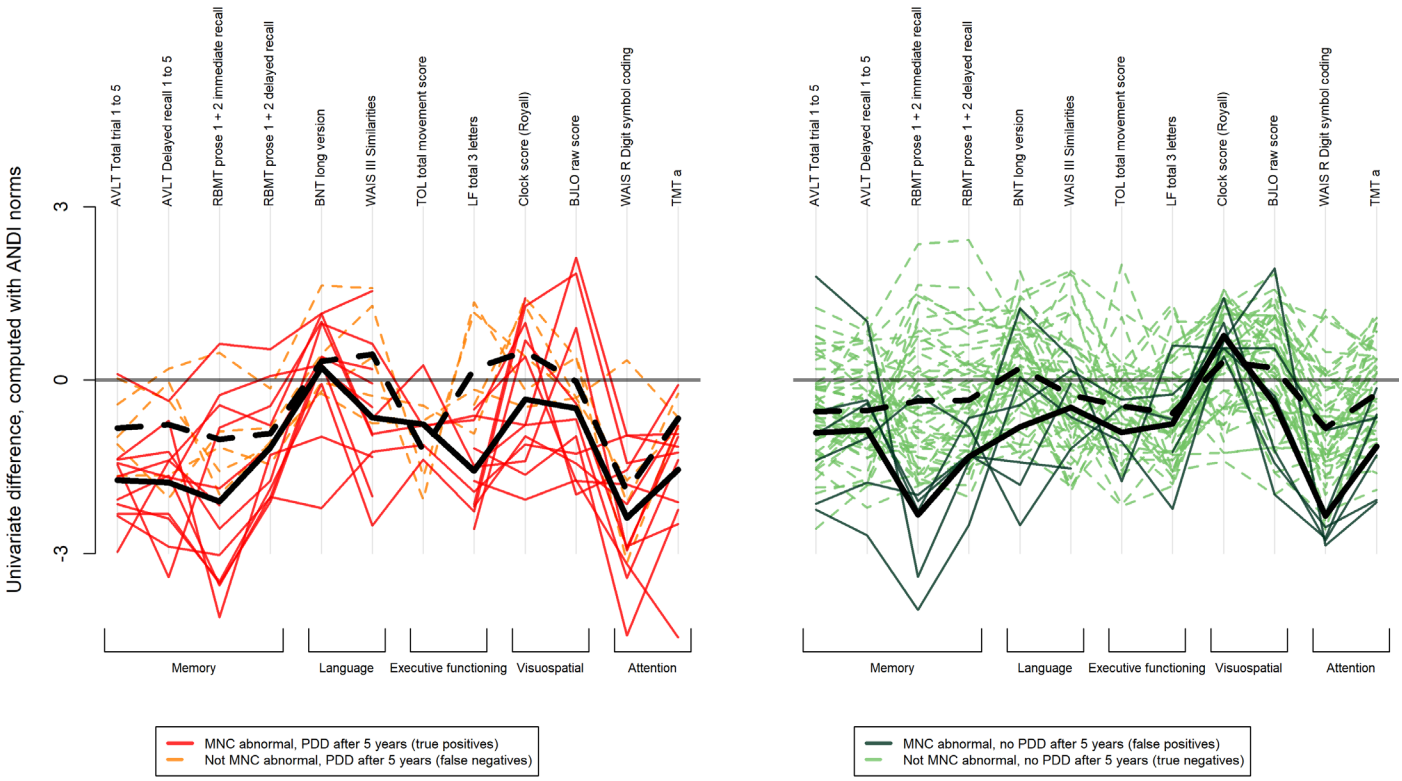


Figure 1. Score profiles of individual patients, in terms of differences between expected scores according to the norm and observed scores. The left panel shows the patients who progressed to PDD after five years. The right panel shows those that did not progress to PDD after five years. Solid lines: patients who are MNC abnormal at baseline, dashed lines: patients who are not MNC abnormal at baseline. The thick black lines

denote the matching mean scores. Note that not all patients completed all tests. Therefore, some lines are interrupted.

7.1.2 Chapter 3 supplement 2: 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. If there were differences, we examined whether the differences between methods in who was diagnosed as impaired, can explain the differences in the performance of the methods in terms of prediction of progression PDD after three and five years.

As can be seen in Table 5, 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 patients who were only diagnosed as impaired by the MNC method. Of the eight patients who were impaired according to the MNC method while they were not impaired according to the traditional PD-MCI criteria, 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%, $\kappa = 0.49$).

Table 5: *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 abnormal			ANDI MNC normal	
		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

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%, $\kappa = 0.63$). Nine patients had PD-MCI according to the criteria applied with ANDI but are normal according to the MNC method. None of these patients progressed to 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 progressed to PDD after three years, and two more patients (four in total) had progressed to dementia after five years. Again, the MNC method identified some patients who would progress to PDD but who were not detected by the PD-MCI method.

Table 6: *Cross-classification table of the two methods applied with ANDI*

		ANDI MNC abnormal			ANDI MNC normal	
		3 years	5 years		3 years	5 years
PD-MCI	24	6 PDD	8 PDD	9	0 PDD	0 PDD
		12 no PDD	4 no PDD		8 no PDD	4 no PDD
		6 missing	12 missing		1 missing	5 missing
no PD-MCI	8	2 PDD	4 PDD	82	1 PDD	5 PDD
		4 no PDD	2 no PDD		64 no PDD	45 no PDD
		2 missing	2 missing		17 missing	31 missing

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%, $\kappa = 0.68$). 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 7 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 normal 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 progressed to PDD who was missed by the traditional method.

Table 7: *Cross-classification table of the two methods using the PD-MCI criteria.*

		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 no PDD
		5 missing	15 missing		3 missing	4 missing
no PD-MCI	3	1 PDD	1 PDD	77	1 PDD	6 PDD
		40no PDD	0 no PDD		60 no PDD	42 no PDD
		2 missing	2 missing		16 missing	29 missing

7.2 Supplements to chapter 4

7.2.1 Chapter 4 supplement 1: Demographic characteristics of each method

Table 1: *Characteristics of the 10 neuropsychological test variables in ANDI that were used for the current analyses.*

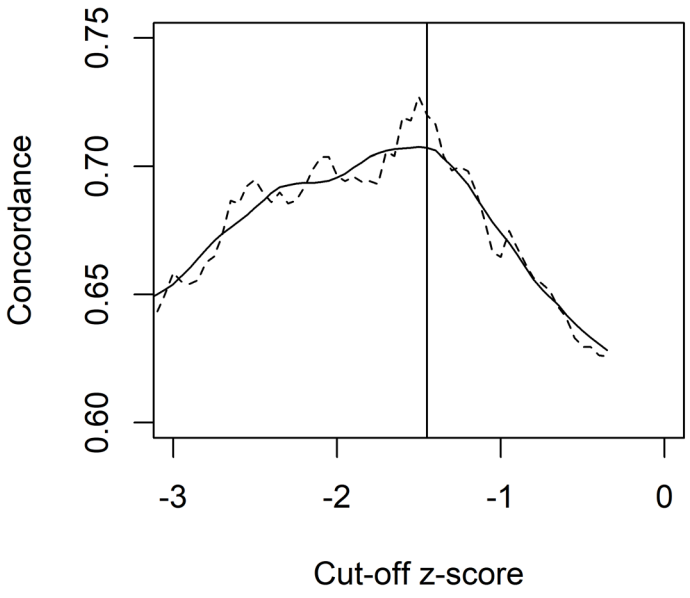
	N in ANDI norms	% Male	Age range	Demographic variables
MMSE	16128	55	18-97	A+S+E
RAVLT – total score of 5 trials	5017	50	18-97	A+S+E
RAVLT – delayed recall score	4540	49	18-97	A+S+E
Animal fluency (1 minute)	5783	40	18-96	A+E
COWAT letter fluency	2894	48	18-97	A+S+E
TMTa	3216	47	18-97	A+S+E
TMTb	3320	46	18-97	A+S+E
Stroop card I – color	1783	41	18–91	A+S+E
Stroop card II – word	2147	43	18-91	A+S+E
Stroop card III – color-word interference	2132	43	18-91	A+S+E

Abbreviations: MMSE=Mini-Mental State Examination; RAVLT=Rey Auditory Verbal Learning Test; COWAT=Controlled Oral Word Association Test; TMT=Trail Making Test; Stroop= Stroop Color-Word Test. A= Age; S= Sex; E= Level of education.

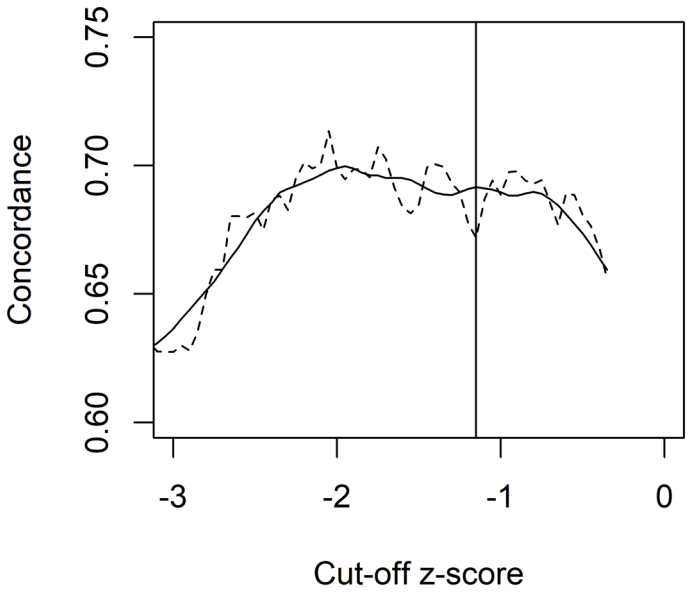
7.2.2 Chapter 4 supplement 2: Calibration of number and magnitude of deviations

For one deviation, visual inspection of Figure 5a suggests that the concordance increases until it reaches a maximum level around z = -1.5, after which it plateaus until about z= -2.5. In this plateau the cut-off scores all produce similar results in terms of predicting progression to dementia. A formal analysis (see supplement 3) was performed to find the border of the plateau which confirms that z = -1.45 is indeed the point where the test statistic starts to plateau. For two deviations (Figure 5b), a similar situation occurs: the test statistic increases until around z = -1 and then has a plateau to z = -2.2; a formal analysis indicates the plateau begins at z=-1.15. For three deviations (Figure 5c), it peaks at z = -0.7 and plateaus until about -2.0; a formal analysis confirms the plateau begins at z=-0.70.

A



B



C

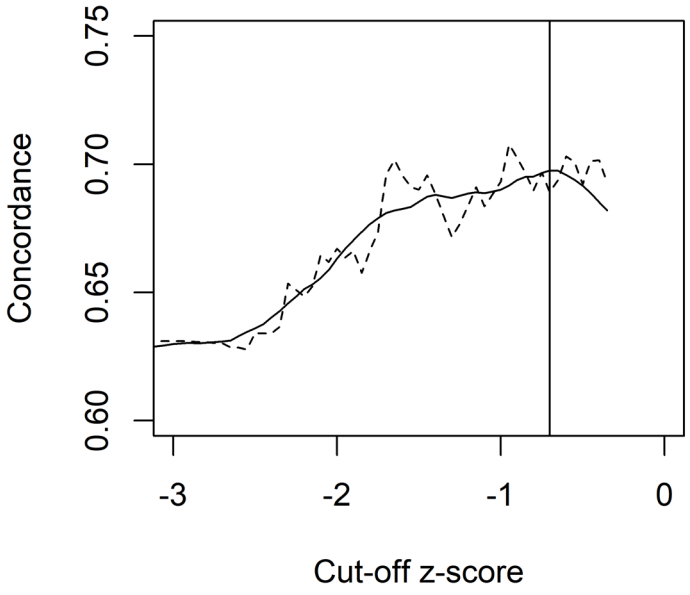


Figure 5 a b c. Concordance (y-axis) as a function of magnitude of cut-off z-scores (x-axis) for prediction of development of dementia given (a) one, (b) two and (c) three deviations. Both the actual concordance scores (dashed line) and a smoothed function (running mean +/- 6; solid line) are plotted. The vertical line indicates the cut-off point found by the formal analysis of the start of the plateau.

For the MNC status we calibrated how large the profile abnormality had to be to best predict progression to dementia. Figure 3 shows how well the MNC-status predicts progression to dementia as a function of the cut-off *p*-values. It shows a clear peak indicating the optimal cut-off value. A *p*-value of 0.22 (one-sided; as we are trying to diagnose abnormality) has to be selected as the best value to predict progression to dementia.

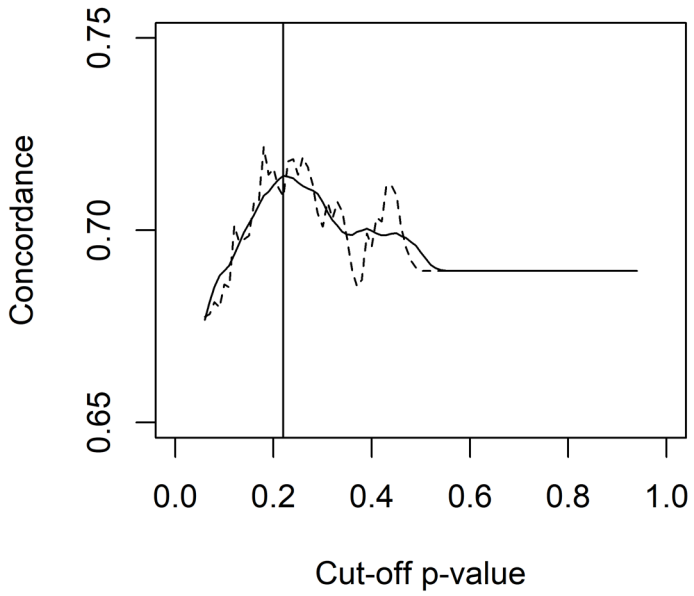


Figure 3. Cut-off *p*-values (x-axis) for the MNC-status method. The y-axis shows the concordance. The dotted line represents the raw curve and the continuous line is the smoothed function (running mean +/- 6). The vertical line indicates the point where the concordance is highest.

7.2.3 Chapter 4 supplement 3: Finding the start of the plateau

In order to find the start of the plateau for each of the criteria (1, 2 or 3 deviations) depicted in Figure 7, we calculated the second derivative. If the second derivative is maximally negative, the concordance has reached its plateau (when reading from right to left). The first derivative was calculated according to:

$$\frac{dy_i}{dx} = \frac{y_{i+2} - y_{i-2}}{x_{i+2} - x_{i-2}}$$

Where y_i is the i^{th} concordance and x_i is i^{th} deviation size. The first derivative denotes the slope (positive means increasing concordance, negative means decreasing predictive ability) and is plotted in Figure 8. Then the second derivative was calculated according to:

$$\frac{d^2y_i}{dx^2} = \frac{dy_{i+2}/dx_{i+2} - dy_{i-2}/dx_{i-2}}{x_{i+2} - x_{i-2}}$$

This denotes the change in slope of the predictor value. This is plotted in Figure 9.

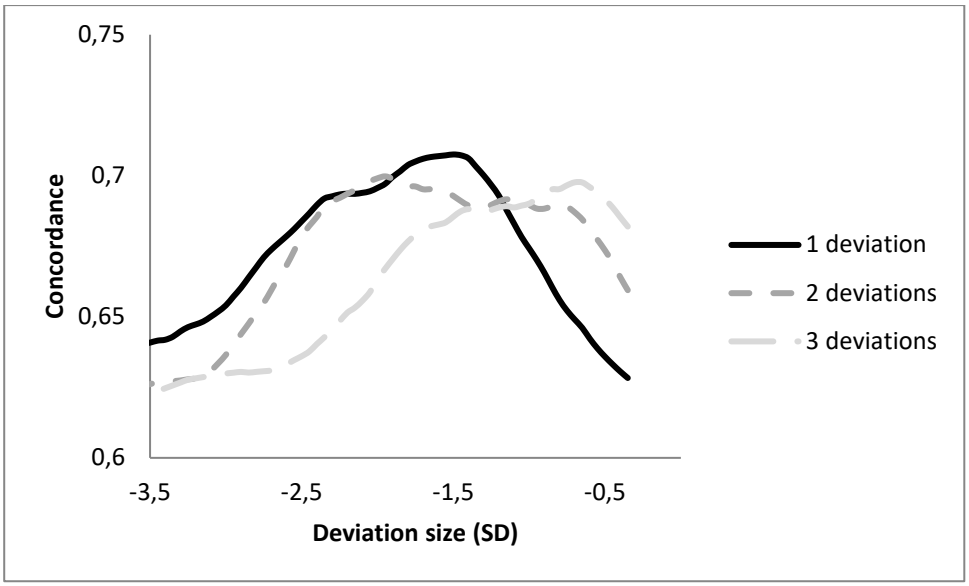


Figure 7. Concordance as a function of deviation size for one, two and three deviations

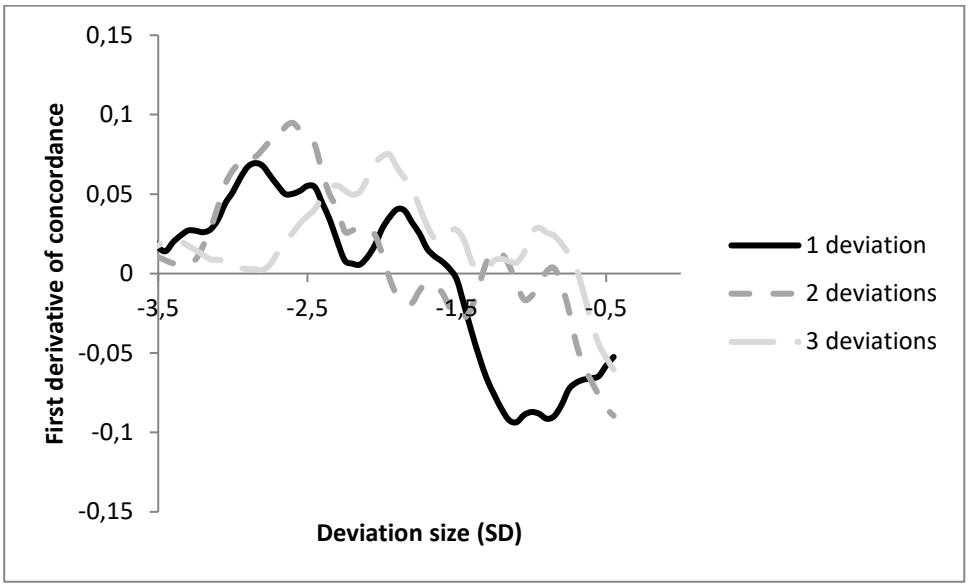


Figure 8. First derivative of concordance as a function of deviation size for one, two and three deviations

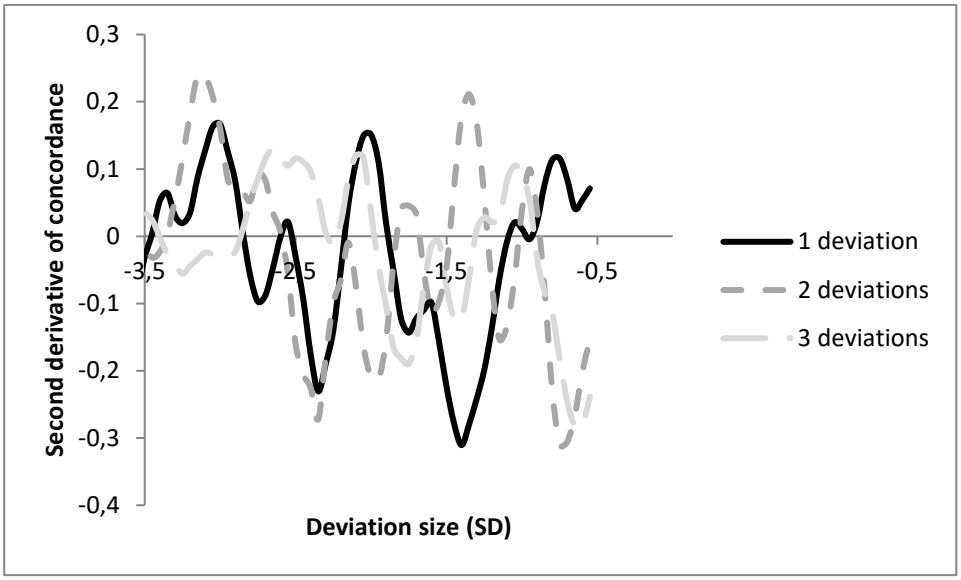


Figure 9. Second derivative of concordance as a function of deviation size for one, two and three deviations. We used this to determine the optimal cut-off points, that is for one deviation it is -1.45, for two -1.15 and for three deviations -0.70.

7.2.4. Chapter 4 supplement 4: Diagnostic overlap between methods.

Table 4: *Diagnostic overlap expressed by Cohen's k for each combination of cut off scores.*

Combination	Cohen's k	95% CI	p
1 deviation (z < -1.45) and 2 deviations (z < -1.15)	.69	.63 - .77	<.001
1 deviation (z < -1.45) and 3 deviations (z < -0.70)	.66	.59 - .74	<.001
2 deviations (z <-1.15) and 3 deviations (z < -0.70)	.68	.61 - .75	<.001
MNC status (p < 0.22) and 1 deviation (z < - 1.45)	.61	.54 - .68	<.001
MNC status (p < 0.22) and 2 deviations (z < -1.15)	.62	.54 - .69	<.001
MNC status (p < 0.22) and 3 deviations (z < -0.70)	.48	.40 - .56	<.001

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. doi:10.1212/WNL.56.6.730.

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. (2018). Multivariate normative comparisons for neuropsychological assessment by a multilevel factor .structure or multiple imputation approach. *Psychological Assessment, 30(4), 436.*

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119-126.

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., ... & Snyder, P. J. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, *7*(3), 270-279.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1993). The spot-the-word test: a robust estimate of verbal intelligence based on lexical decisions. *British Journal of Clinical Psychology, 32*(1), 55-65.

Barona, A., Reynolds, C. R., & Chastain, R. (1984). A demographically based index of premorbid intelligence for the WAIS-R. *Journal of Consulting and Clinical Psychology, 52*(5)*,* 885-887.

Basso, M.R., Bornstein, R.A., Roper, B.L., & McCoy, V.L. (2000). Limited accuracy of premorbid intelligence estimators: a demonstration of regression to the mean. *The Clinical Neuropsychologist, 14(3),* 325-340.

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.

Van Den Berg, E., Dekker, J. M., Nijpels, G., Kessels, R. P., Kappelle, L. J., De Haan, E. H., ... & Biessels, G. J. (2008). Cognitive functioning in elderly persons with type 2 diabetes and metabolic syndrome: the Hoorn study. *Dementia and geriatric cognitive disorders*, *26*(3), 261-269.

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.

Blair, J.R., & Spreen, O. (1989). Predicting Premorbid IQ: A Revision of the National Adult Reading Test. *The Clinical Neuropsychologist, 3(2),* 129-136.

Bondi, M. W., Edmonds, E. C., Jak, A. J., Clark, L. R., Delano-Wood, L., McDonald, C. R., ... & Salmon, D. P. (2014). Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. *Journal of Alzheimer's Disease*, *42*(1), 275-289.

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc. Ser. B*, 26, 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.

Carley, S., Dosman, S., Jones, S. R., & Harrison, M. (2005). Simple nomograms to calculate sample size in diagnostic studies. *Emergency Medicine Journal*, *22*(3), 180-181.

Castelli, L., Rizzi, L., Zibetti, M., Angrisano, S., Lanotte, M., and Lopiano, L. (2010). Neuropsychological changes 1-year after subthalamic DBS in PD patients: a prospective controlled study. *Parkinsonism Relat. Disord.* 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.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge.

Cohen, S., Ter Stege, J. A., Geurtsen, G. J., Scherpbier, H. J., Kuijpers, T. W., Reiss, P., et al. (2015). Poorer cognitive performance in perinatally HIV-infected children versus healthy socioeconomically matched controls. *Clin. Infect. Dis.* 60, 1111–1119.

Collin, C., Wade, D. T., Davies, S., & Horne, V. (1988). The Barthel ADL Index: A reliability study. *International Disability Studies*, *10*(2), 61-63.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-202.

Crawford, J. R., & Allan, K. M. (1997). Estimating premorbid WAIS-R IQ with demographic variables: Regression equations derived from a UK sample. *The Clinical Neuropsychologist*, *11*(2), 192-197.

Crawford, J. R., and 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., and Laws, K. R. (2006). Testing for a deficit in single-case studies: effects of departures from normality. *Neuropsychologia* 44, 666–677.

Crawford, J. R., and Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. *Clin. Neuropsychol.* 12, 482–486.

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., and Hussong, A. M. (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychol. Methods* 14, 81.

Dennis, M., Francis, D. J., Cirino, P. T., Schachar, R., Barnes, M. A., & Fletcher, J. M. (2009). Why IQ is not a covariate in cognitive studies of neurodevelopmental disorders. *Journal of the International Neuropsychological Society*, *15,* 311-343.

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, B., Burn, D., Goetz, C., Aarsland, D., Brown, R. G., Broe, G. A., ... & Korczyn, A. (2007). Diagnostic procedures for Parkinson's disease dementia: recommendations from the movement disorder society task force. *Movement disorders*, *22*(16), 2314-2324.

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.

Flicker C, Ferris SH, Reisberg B. Mild cognitive impairment in the elderly: predictor of dementia. *Neurology* 1991; 41: 1006–09.

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.

Van Der Flier, W. M., & Scheltens, P. (2018). Amsterdam dementia cohort: performing research to optimize care. *Journal of Alzheimer's Disease*, *62*(3), 1091-1111.

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.

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.

Gershon, R. C., Cella, D., Fox, N. A., Havlik, R. J., Hendrie, H. C., & Wagster, M. V. (2010). Assessment of neurological and behavioural function: the NIH Toolbox. *The Lancet Neurology*.

González-Redondo, R., Toledo, J., Clavero, P., Lamet, I., García-García, D., García-Eulate, R., et al. (2012). The impact of silent vascular brain burden in cognitive impairment in Parkinson's disease. *Eur. J. Neurol.* 19, 1100–1107.

Grober, E., Sliwinsk, M., & Korey, S. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology, 13(6),* 933-949.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, *15*(4), 361-387.

Hobson, P., & Meara, J. (2004). The risk and incidence of dementia in a cohort of older subjects with Parkinson's disease in 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.

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.

Huizenga, H. M., Agelink van Rentergem, J. A., Grasman, R. P., Muslimovic, D., and Schmand, B. (2016). Normative comparisons for large neuropsychological test batteries: user-friendly and sensitive solutions to minimize familywise false positives. *J. Clin. Exp. Neuropsychol.* 38, 611–629.

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

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J.-M., and Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Comput. Stat. Data Anal.* 51, 5142–5154.

Janssen, M. A. M., Bosch, M., Koopmans, P. P., & Kessels, R. P. C. (2015). Validity of the Montreal Cognitive Assessment and the HIV Dementia Scale in the assessment of cognitive impairment in HIV-1 infected patients. *Journal of neurovirology*, *21*(4), 383-390.

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.

Jutten, R. J., Harrison, J., de Jong, F. J., Aleman, A., Ritchie, C. W., Scheltens, P., & Sikkes, S. A. (2017). A composite measure of cognitive and functional progression in Alzheimer's disease: Design of the Capturing Changes in Cognition study. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *3*(1), 130-138.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *Boston Naming Test.* Philadelphia, PA: Lea & Febiger. Kobayakawa, M., Koyama, S., Mimura, M., & Kawamura, M. (2008). Decision making in Parkinson's disease: Analysis of behavioral and physiological patterns in the Iowa gambling task. *Movement disorders*, *23*(4), 547-552.

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

Kramer, J. H., Mungas, D., Possin, K. L., Rankin, K. P., Boxer, A. L., Rosen, H. J., ... & Widmeyer, M. (2014). NIH EXAMINER: conceptualization and development of an executive function battery. *Journal of the international neuropsychological society*, *20*(1), 11-19.

Leritz, E.C., McGlinchey, R.E., Lundgren, K., Grande, L.J., & Milberg, W.P. (2008). Using Lexical Familiarity Judgements to Assess Verbally-Mediated Initelligence in Aphasia. *Neuropsychology, 22(6),* 687-696.

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. *J. Exp. Soc. Psychol.* 49, 764–766.

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

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.

Matarazzo, J. D. (1972). *Wechsler's Measurement and Appraisal of Adult Intelligence* (5thed.). New York: Oxford University Press.

Mattsson, N., Zetterberg, H., Hansson, O., Andreasen, N., Parnetti, L., Jonsson, M., ... & Rich, K. (2009). CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. *Jama*, *302*(4), 385-393.

McKeith, I. G., Dickson, D. W., Lowe, J., Emre, M., O'brien, J. T., Feldman, H., ... & Aarsland, D. (2005). Diagnosis and management of dementia with Lewy bodies third report of the DLB consortium. *Neurology*, *65*(12), 1863-1872.

Muslimović, D., Post, B., Speelman, J. D., & Schmand, B. (2005). Cognitive profile of patients with newly diagnosed Parkinson disease. *Neurology*, *65*(8), 1239-1245. doi: 10.1212/01.wnl.0000180516.69442.95.

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*(03), 426-437.

Neary, D., Snowden, J. S., Gustafson, L., Passant, U., Stuss, D., Black, S. A. S. A., ... & Boone, K. (1998). Frontotemporal lobar degeneration A consensus on clinical diagnostic criteria. *Neurology*, *51*(6), 1546-1554.

Nelson, H. E. (1976). A modified card sorting test sensitive to frontal lobe defects. *Cortex*, *12*(4), 313 324. Pagonabarraga, J., Kulisevsky, J., Llebaria, G., García-Sánchez, C., Pascual-Sedano, B., & Gironell, A. (2008). Parkinson's disease-cognitive rating scale: A new cognitive scale specific for Parkinson's disease. *Movement Disorders*, *23*(7), 998-1005.

Nuechterlein, K. H., Green, M. F., Kern, R. S., Baade, L. E., Barch, D. M., Cohen, J. D., ... & Goldberg, T. (2008). The MATRICS Consensus Cognitive Battery, part 1: test selection, reliability, and validity. *American Journal of Psychiatry*, *165*(2), 203-213.

Olsson, B., Lautner, R., Andreasson, U., Öhrfelt, A., Portelius, E., Bjerke, M., ... & Wu, E. (2016). CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *The Lancet Neurology*, *15*(7), 673-684.

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.

Petersen, R.C., Smith G.E., Waring S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Arch Neurol*, 56(3), 303-308.

Petersen, R. C., & Morris, J. C. (2005). Mild cognitive impairment as a clinical entity and treatment target. *Archives of neurology*, *62*(7), 1160-1163.

Petersen, R. C. (2016). Mild cognitive impairment. *CONTINUUM: Lifelong Learning in Neurology*, *22*(2 Dementia), 404.

Van der Putten, J. J., Hobart, J. C., Freeman, J. A., & Thompson, A. J. (1999). Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel index and the Functional Independence Measure. *Journal of Neurology, Neurosurgery & Psychiatry*, *66*(4), 480-484.

Rapp, S., Brenes, G. M. A. P., & Marsh, A. P. (2002). Memory enhancement training for older adults with mild cognitive impairment: a preliminary study.

Rascovsky, K., Hodges, J. R., Knopman, D., Mendez, M. F., Kramer, J. H., Neuhaus, J., ... & Hillis, A. E. (2011). Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain*, *134*(9), 2456-2477.

Reitan, R.M. (1992). *Trail Making Test: Manual for administration and scoring.* Tucson, AZ: Reitan Neuropsychological Laboratory.

Rey, A. (1958). *L'Examen Clinique en Psychologie.* Paris: Presses Universitaries de France. Royall, D.R., Cordes, D.A., & Polk, M. (1998). CLOX: An executive clock drawing task. *Journal of Neurology, Neurosurgery, and Psychiatry, 64,* 588-594.

Román, G. C., Tatemichi, T. K., Erkinjuntti, T., Cummings, J. L., Masdeu, J. C., Garcia, J. A., ... & Moody, D. M. (1993). Vascular dementia Diagnostic criteria for research studies: report of the NINDS-AIREN International Workshop. *Neurology*, *43*(2), 250-250.

Ruan, Q., D'Onofrio, G., Sancarlo, D., Bao, Z., Greco, A., & Yu, Z. (2016). Potential neuroimaging biomarkers of pathologic brain changes in Mild Cognitive Impairment and Alzheimer's disease: a systematic review. *BMC geriatrics*, *16*(1), 104.

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *Statistician*, 41, 169–178. doi: 10.2307/2348250

Salthouse, T. A. (2000). Aging and measures of processing speed. *Biological psychology*, *54*(1-3), 35-54.

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.

Schoenberg, M.R., Duff, K., Scott, J.G., Patton, D., & Adams, R.L. (2006).Prediction errors of the Oklahoma Premorbid Intelligence Estimate-3 (OPIE-3) stratified by 13 age groups. *Archives of clinical neuropsychology, 21*(5), 469-475.

Schmand, B., De Bruin, E., De Gans, J., and van de Beek, D. (2010). Cognitive functioning and quality of life nine years after bacterial meningitis. *J. Infect.* 61, 330–334.

Schwab, J.F., & England, A.C., (1969). Projection technique for evaluating surgery in Parkinson's disease. In: Billingham, F.H., Donaldson, M.C. (Eds.), Third Symposium on Parkinson's Disease. Churchill Livingstone, Edinburgh, pp. 152–157.

Smeding, H. M., Speelman, J. D., Huizenga, H. M., Schuurman, P. R., and Schmand, B. (2011). Predictors of cognitive and psychosocial outcome after STN DBS in Parkinson's Disease. *J. Neurol. Neurosurg. Psychiatry* 82, 754–760.

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

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.

Stouten-Kemperman, M. M., de Ruiter, M. B., Koppelmans, V., Boogerd, W., Reneman, L., & Schagen, S. B. (2015). Neurotoxicity in breast cancer survivors≥ 10 years post-treatment is dependent on treatment type. *Brain imaging and behavior*, *9*(2), 275-284.

Su, T., Schouten, J., Geurtsen, G. J., Wit, F. W., Stolte, I. G., Prins, M., et al. (2015). Multivariate normative comparison, a novel method for more reliably detecting cognitive impairment in HIV infection. *AIDS* 29, 547–557.

Testa, S. M., Winicki, J. M., Pearlson, G. D., Gordon, B., and Schretlen, D. J. (2009). Accounting for estimated IQ in neuropsychological test performance with regression-based techniques. *J. Int. Neuropsychol. Soc.* 15, 1012–1022.

UNESCO. (1997). *International Standard Classification of Education-ISCED 1997: November 1997*. UNESCO.

UNESCO (2011). *International Standard Classification of Education-ISCED 2011: December 2012.* UNESCO.

Valdés-Sosa, M., Bobes, M. A., Quiñones, I., Garcia, L., Valdes-Hernandez, P. A., Iturria, Y., et al. (2011). Covert face recognition without the fusiform-temporal pathways. *Neuroimage* 57, 1162–1176.

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.

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.

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*, Van Gorcum Prakke en Prakke, Assen.

Virk, G. K., Poljak, A., Braidy, N., & Sachdev, P. S. (2018). CSF and blood biomarkers of early-onset Alzheimer's Disease: A systematic review and meta-analysis. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *14*(7), 1158.

Whittle, C., Corrada, M. M., Dick, M., Ziegler, R., Kahle-Wrobleski, K., Paganini-Hill, A., et al. (2007). Neuropsychological data in nondemented oldest old: the 90+ study. *J. Clin. Exp. Neuropsychol.* 29, 290–299.

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.

Wilkinson, G.S., & Robertson, G. J. (2006). Wide Range Achievement Test (WRAT4). *Psychological Assessment Resources, Lutz.*

Wilson, B., Cockburn J., & Baddeley, A. (1983). *Rivermead Behavioural Memory Test.* Reading, UK: Thames Valley Test Company.

Zachary, R. A., & Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of Clinical psychology*, *41*(1), 86-94.

Funding, author contributions, and short CV

Funding Information

The studies in this dissertation were all supported by a grant from the Netherlands Organization for Scientific Research (NWO) awarded to prof. dr. B.A. Schmand, prof. dr. J.M.J. Murre, and H.M. Huizenga under grant 480-12-015. The University of Amsterdam provided additional funding for this project.

Author Contributions

Chapter 2

N.R. de Vent*, J.A. Agelink van Rentergem*, B.A. Schmand, J.M.J. Murre, ANDI Consortium, & H.M. Huizenga, (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 and revised the manuscript. JAAvR developed the method, analysed the data, wrote and revised the manuscript. BS developed the concept, supervised data collection and assembling the database, revised the manuscript, JM developed the concept, revised the article. ANDI consortium donated the data for the database. HH developed the concept, supervised the analysis and revised the manuscript.

Chapter 3

J.A. Agelink van Rentergem*, N.R. de Vent*, H.M. Huizenga, J.M.J. Murre, ANDI Consortium, & B.A. Schmand, (2019). Predicting Progression to Parkinson's Disease Dementia using Modern Neuropsychological Techniques. *Journal of the International Neuropsychological Society*, 25(7), 678-687

JAAvR developed the method, analysed the data, wrote and revised the manuscript. NRdV developed the method, assembled the database, wrote and revised the article. HH supervised the analysis, revised the article. JM revised the article. ANDI consortium provided the data. BS developed the concept, provided data, supervised data collection and database assembly, revised the article.

Chapter 4

N.R. de Vent, H.M. Huizenga, J.M.J. Murre, J.A. Agelink van Rentergem, W.M. van der Flier, S.A.M., Sikkes,K van der Bosch, ANDI Consortium, & B.A. Schmand (2020). An Operational Definition Of 'Abnormal Cognition' To Optimally Predict Progression To Dementia. What are optimal cut-off points for univariate and multivariate normative comparisons? *Journal of Alzheimers's Disease, 1-11*

NRdV developed the concept, developed the method, assembled the database, analysed the data, wrote and revised the manuscript. HH supervised the data analysis and revised the manuscript. JM revised the manuscript. JAAvR analysed the data and revised the manuscript. WMvdF developed the concept, donated data, revised the manuscript SAMS revised the manuscript. KvdB revised the manuscript. ANDI consortium donated the data for the database. BS developed the concept, developed the method, revised the manuscript.

Chapter 5

N.R. de Vent, J.A. Agelink van Rentergem, M.C. Kerkmeer, H.M. Huizenga, B.A. Schmand,& J.M.J. Murre (2016). Universal Scale of Intelligence Estimates (USIE): representing intelligence estimated from level of education. *Assessment, 25*(5), 557-563

NRdV developed the concept, developed the method, analysed the data, wrote and revised the manuscript. JAavR developed the concept and revised the manuscript MCK donated data and revised the manuscript. HH developed the concept and revised the manuscript. JM developed the concept, developed the method, supervised the analyses and revised the manuscript.

Other publications

Putkinen, V., Tervaniemi, M., Saarikivi, K., de Vent, N., & Huotilainen, M. (2014). Investigating the effects of musical training on functional brain development with a novel Melodic MMN paradigm. *Neurobiology of learning and memory*, *110*, 8-15.

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.*

Nathalie Ramona de Vent was born in Hoorn, on the 9th of July 1987. After C.B. de Rank and Oscar Romero she completed a propaedeutic year of Social Work at the Amsterdam University of Applied Science after which she studied Educational Sciences at the University of Amsterdam and obtained her bachelor's degree in 2010. She completed a two years Research Master in Cognitive Neuropsychology at VU University with an internship at the University of Helsinki. The research for this PhD project was conducted between September 2013 and March 2020 and was supervised by prof.dr. Ben Schmand, prof. dr. Jaap Murre, and prof.dr. Hilde Huizenga. The research in this thesis was conducted in close collaboration with Joost Agelink van Rentergem.

Nederlandse samenvatting

Het doel van dit proefschrift was om problemen bij de neuropsychologische diagnostiek van patiënten te verhelpen en om het diagnostisch proces verder te verbeteren. We hebben ons voornamelijk gericht op het verbeteren van normatieve vergelijkingen die worden uitgevoerd om de testscores van patiënten te beoordelen en om te beslissen of een score normaal of abnormaal is. Ten eerste hebben we een grote normatieve database gemaakt, de Advanced Neuropsychological Diagnostics Infrastructure (ANDI), die correcties voor leeftijd, geslacht en opleidingsniveau mogelijk maakt. Op deze manier kan elke patiënt worden vergeleken met een relevante set van gezonde controledeelnemers. Ten tweede hebben we een multivariate normatieve vergelijkingsmethode toegepast. Deze methode evalueert het profiel van de scores die een patiënt heeft behaald.

Hoofdstuk 2 beschreef de constructie van de geaggregeerde database. Door gedoneerde datasets van gezonde deelnemers te combineren, konden we een normatieve database opbouwen zonder de deelnemers zelf te testen. Data van gezonde deelnemers kwamen van controlepersonen in klinische onderzoeken en grote populatie studies. Om uit meerdere datasets een grote, samenhangende database te maken, is een gestandaardiseerde procedure gevolgd die uit vijf stappen bestond.

(1) Extreme grenzen. Om ANDI gevoelig te maken voor het detecteren van abnormale prestaties zijn er scores verwijderd die deze detectiecapaciteit in gevaar zouden kunnen brengen. We hebben dit in twee stappen gedaan. Allereerst hebben we gecontroleerd of scores van deelnemers behaald waren binnen de grenzen van de test. Als scores onmogelijk hoog bleken (bijv. hoger dan de perfecte scores), werden deze verwijderd. Met behulp van de klinische expertise van onze ANDI-stuurgroep hebben we ook voor iedere test een ondergrens gedefinieerd. Als deelnemers onder deze grens scoorden, konden ze niet langer als cognitief gezond worden bestempeld en werden hun scores ook verwijderd.

(2) Modelselectie: Om te bepalen wat de invloed van de verschillende demografische kenmerken is, hebben we regressie analyses gedaan moesten er modellen gefit worden. We hebben het best passende regressiemodel gebaseerd op het Akaike Information Criterion (AIC).

(3) Uitbijters verwijderen. Op basis van de toegepaste modellen hebben we demografisch gecorrigeerde uitbijters, dat wil zeggen scores die erg laag waren gezien het geslacht, de leeftijd en het opleidingsniveau van een deelnemer, verwijderd.

(4) Normaliteit. Om parametrische tests te kunnen gebruiken en omdat veel van de neuropsychologische testdata niet normaal verdeeld waren, hebben we deze scores omgezet in normaal verdeelde scores. Voor iedere test variabele werd de Box-Cox-procedure (Box & Cox, 1964) gebruikt om de best mogelijke transformatie te vinden. Deze transformatie werd vervolgens toegepast op de data.

(5) Modellen opnieuw fitten. Met de opgeschoonde en genormaliseerde data is berekend welke regressiegewichten voor de demografische variabelen we hebben moesten implementeren op de ANDI-website. Hier is dus gekeken welke invloed van leeftijd, sekse, en opleidingsniveau voor iedere test variabele relevant zijn. Hiervoor hebben we weer de AIC gebruikt.

Het laatste deel van hoofdstuk 2 beschreef de inhoud van de ANDI-database. Voor elke test beschreven we het aantal gedoneerde datasets, het aantal deelnemers dat de betreffende test had gedaan, het leeftijdsbereik, het percentage mannen en vrouwen, en het opleidingsbereik. Bijvoorbeeld, de auditieve verbale leertest (AVLT; testvariabele uitgestelde reproductie) was gedoneerd door 29 studies. Deze donatie bevatte gegevens van 4598 gezonde individuen, van wie de leeftijd varieerde van 14 tot 97 jaar. Van hen waren 49% mannen en het onderwijsniveau varieerde van 1 (onafgemaakt onderwijs) tot 7 (universitair diploma)).

In hoofdstuk 3 werd onze aanpak om de ANDI-database te combineren met de multivariate normatieve vergelijkingsmethode getest door longitudinale data van patiënten de ziekte van Parkinson (PD) opnieuw te analyseren. Een deel van deze patiënten ontwikkelde op een later moment dementie (PDD) (Broeders). et al., 2013)). Eerder werden conventionele criteria voor de ziekte van Parkinson met milde cognitieve stoornissen (PD-MCI; Litvan et al., 2012) gebruikt om de progressie naar PDD te voorspellen (Muslimovic et al., 2005; Muslimovic et al, 2009; Broeders et al, 2013). In het huidige hoofdstuk hebben we onderzocht of het combineren van ANDI en de multivariate normatieve vergelijkingen de voorspelling van progressie naar PDD zouden verbeteren. Eerst keken we naar de PD-MCI-criteria op de conventionele

(univariate) manier, maar gebruik makend van de normatieve scores van de ANDI-database. Dat wil zeggen, we hebben de PD-MCI-criteria toegepast, maar in plaats van traditionele normatieve data te gebruiken, gebruikten we de ANDI-database en de demografisch gecorrigeerde normen. Dit toonde aan dat minder patiënten op basis van de ANDI-database als abnormaal werden bestempeld dan met de traditionele normen. De gevoeligheid en specificiteit van de PD-MCI-criteria met traditionele normen waren vergelijkbaar met de PD-MCI criteria met ANDI-normen. Dat wil zeggen dat het gebruik van de ANDI-database op zich de voorspelling niet verbeterde. Daarnaast hebben we gekeken of het gebruik van de multivariate normatieve vergelijkingen samen met de ANDI-database zou resulteren in een betere voorspelling van welke patiënten later dementie zouden krijgen ten opzichte van de conventionele PD-MCI-criteria. Dit bleek het geval te zijn. Zowel gevoeligheid als specificiteit waren hoger voor de multivariate normatieve vergelijkingsmethode dan voor de conventionele PD-MCI-criteria (met traditionele normen of met de ANDI-database). De resultaten van dit onderzoek gaven aan dat e multivariate normatieve vergelijkingen met de ANDI-database een meer gevoelige voorspelling van PDD mogelijk maakten dan de traditionele PD-MCI-criteria.

In hoofdstuk 4 werden ANDI en de multivariate vergelijkingsmethode gebruikt om 'abnormale cognitie' te helpen definiëren in een groep niet-demente patiënten van het Amsterdamse dementiecohort (van der Flier & Scheltens, 2018). In de literatuur zijn verschillende definities van abnormale cognitie gegeven, maar er is nog geen consensus. In dit hoofdstuk hebben we onderzocht hoe abnormale cognitie het beste kan worden gedefinieerd bij het voorspellen van progressie naar dementie, en of de multivariate normatieve vergelijkingsmethode deze voorspelling zou kunnen verbeteren. Ten eerste hebben we met behulp van een kruisvalidatiemethode bepaald welk aantal abnormale tests en welke grootte van scoreafwijkingen de progressie naar dementie het beste voorspelden. Met behulp van 10 neuropsychologische testscores ontdekten we dat het vermogen om progressie naar dementie te voorspellen op basis van één, twee en drie abnormale testscores vergelijkbaar is, op voorwaarde dat de afkapwaarden op de juiste manier worden aangepast. Dat wil zeggen, wanneer één afwijkende score vereist is, moet de afwijking groter zijn dan wanneer drie afwijkende scores vereist zijn. We ontdekten ook dat een multivariate profielanalyse aanvullende informatie geeft over de cognitie van een persoon, zelfs als hij/zij

voldoet aan de conventionele MCI-criteria. Deze profielanalyse kan dus de voorspelling van progressie naar dementie verder verbeteren.

Hoofdstuk 5 nam een zijstap en keek naar de proxies voor premorbide intelligentie welke worden gebruikt bij neuropsychologische evaluatie. Omdat echte premorbide IQ-scores van patiënten bijna nooit beschikbaar zijn, maken we over het algemeen gebruik van een proxy, bijvoorbeeld tests die minder kwetsbaar zijn voor neuropsychologische problemen. Ook kan een schatting van premorbide intelligentie worden verkregen door te vragen naar de achtergrond van een patiënt zoals werkgeschiedenis of opleidingsniveau. Opleidingsniveau is de meest gebruikte proxy, omdat dit bijna altijd bekend is. Het wordt echter op veel verschillende schalen uitgedrukt, variërend van hoge resolutie tot zeer ruw (laag, gemiddeld, hoog), en dit maakt het moeilijk om de schalen te vergelijken (in bijvoorbeeld meta-analyse setting). We stellen daarom een nieuwe schaal voor het premorbide IQ voor die uitwisselbaar is met alle andere: de Universal Scale of Intelligence Estimates (USIE). De USIE vertaalt verschillende bestaande onderwijsschalen in (pseudo) IQ-scores. Wanneer dit wordt gedaan voor een aantal verschillende onderwijsschalen, kunnen we elke schaal uitdrukken in termen van USIE IQ-scores. Op deze manier kunnen alle onderwijsschalen naar elkaar worden vertaald door de USIE IQ-schaal, die wordt gebruikt als tussenliggende schaal. Als alle onderwijsschalen worden vertaald naar de USIE, kan de USIE meta-analyses en andere soorten onderzoek faciliteren.

Het is van het grootste belang dat neuropsychologen hun patiënten kunnen vergelijken met een representatief gezond normenbestand. Dit houdt in dat er wordt vergeleken in termen van leeftijd, geslacht en opleidingsniveau. Dit bevordert een betrouwbaardere detectie van patiënten die cognitief abnormaal zijn en die een vorm van interventie of zorg nodig hebben. De ANDI-database heeft aangetoond de normatieve vergelijkingen te verbeteren door meer representatieve normen te bieden en door een verbetering van statistische toetsingen mogelijk te maken.

Acknowledgements - Dankwoord

Eerst bedank ik natuurlijk mijn promotoren. Ben, jouw onvermoeibare inzet tot ver na je pensioen hebben mij letterlijk en figuurlijk geholpen me over de eindstreep te trekken. Je ongeëvenaarde ervaring en kennis als clinicus zijn onmisbaar gebleken maar ook voor de ontspanning had je altijd wel een mooi verhaal uit de oudheid. Jaap, na vele overleggen als ANDI-team staan we nu aan het roer van de ANDI-Norms B.V. Ik voorzie nog vele successen in de nabije toekomst. Hilde, copromotor, maar dat voelt voor mij zeker niet zo. Naast je fijne secure feedback blijf je ook oog houden voor de personen achter je promovendi. Ik heb dit altijd als heel prettig ervaren.

De commissieleden, Sanne, Roy en Esther als ANDI-stuurgroep leden nu ook in mijn commissie, heel veel dank. Edward, Joke, en Sieberen ik wil jullie ook heel erg bedanken voor jullie deelname aan de commissie.

Joost, we zijn al een tijdje uit elkaar maar ik mis onze wittebroodsweken nog met regelmaat hoor! Samen hebben we het toch maar geflikt. Hopelijk kunnen we in de toekomst nog eens samen wat oppakken.

Fijne collega's Laura, Tycho, veel gezelligheid op de Diamandbeurs! Jacqueline, Maaike, Daan, de tweede generatie kamergenoten. Het was altijd gezellig met jullie op kantoor. Rooske, sorry dat we een bureau moesten delen. Ik vond het leuker toen we op 1 kamer zaten samen. Jessica, Anne-Wil en Christiane, de derde generatie kamergenoten. Maar zeker niet minder gezellig. Het was fijn jullie frisse moed als nieuwe AIO's te observeren. Ik ben benieuwd hoe het jullie de aankomende jaren zal gaan. Ongetwijfeld prima. Nihayra, bedankt voor je fijne gesprekken en je pas voor de kolfkamer als ik die weer eens kwijt was!

Andere gang-genoten van OP (Brenda, Helle, Ellen, Tjistke, etc.) bedankt voor de gezelligheid op de afdeling. Ik als BC-er heb me altijd erg welkom gevoeld op jullie afdeling. Ongetwijfeld ben ik nog mensen vergeten te noemen, maar ook jullie worden bedankt voor de fijne tijd!

Bart, bedankt dat je mijn paranymf wilde zijn. Het was een logisch gevolg op het zijn van mijn getuigen tijdens mijn trouwerij. Besties for life!

Alantia, hartelijk dank voor de fijne speelsessies tijdens het afronden van dit beestenwerk. Mark, bedankt voor al je inzet in het afronden van Nova Zembla. Het laatste half jaar van mijn proefschrift heb ik letterlijk naar niets anders geluisterd en heeft dat me er doorheen geloodst!

Andere vrienden ook bedankt voor jullie support de afgelopen jaren. Speciale vermelding nog voor Sarah, wie het begin van dit werk niet heeft meegekregen maar ontzettend fijne steun is geweest de afgelopen tijd.

Mijn moeder, vader, broer en schoonzus voor jullie support over (al) die jaren studie. Jullie zijn toppers. Zonder al die kont-schoppen zou dit proefschrift er nooit zijn gekomen.

Mijn schoonouders voor hun eindeloze oppas afspraken zodat ik aan mijn boekje kon werken. Jan en Clara bedankt! Alle andere familieleden ook bedankt natuurlijk!

Erik en de meiden. Dit proefschrift is natuurlijk fantastisch maar hoe ons gezin is gegroeid tijdens deze promotieperiode is natuurlijk altijd nog vele male mooier. Bedankt dat ik zoiets moois heb om na mijn werk thuis te komen. Ik hou van jullie!

Met vriendelijke groet,
Nathalie