Statistical advances in clinical neuropsychology

Agelink van Rentergem Zandvliet, J.A.

Publication date
2018

Document Version
Other version

License
Other

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
5.1 Abstract

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

5.2 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

To summarize, neuropsychology would benefit from clarity on the number and type of cognitive domains, and on which tests belong to which cognitive domains. This would facilitate test selection, diagnosis of single-domain and multi-domain disorders, calculation of composite scores, and neuropsychological research into the construct validity of tests.
5.3 Study 1: Factor Meta-Analysis

5.3.1 Methods

5.3.1.1 Literature search

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

5.3.1.2 Exclusion criteria

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

5.3.1.3 Tests

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

5.3.1.4  Contacting authors

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

The authors were sent a list of variables of interest that were to be included, which were the test variables that they collected in their study, along with age, sex, and level of education. There was no specific hypothesis for the influence of sex on the factor structure, but we chose to correct for its influence as well because this is common in neuropsychology (Testa, Winicki, Pearlson, Gordon, & Schretlen, 2009). Level of education was scored differently in different studies, sometimes using a seven-point-scale, sometimes using years of education. This issue is discussed in more depth in the discussion section and is addressed in study 2. Authors were requested to send a correlation matrix of these variables, for the cognitively healthy sample within their data. If they were unsure that their participants qualified as cognitively healthy, possibilities for exclusion criteria within their data were discussed. For example, if measurements from the Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975) and Clinical Dementia Rating (CDR; Morris, 1997) had been taken in their study, participants with MMSE scores below 24 and CDR scores above 0 could be removed before the correlation matrix was computed. Since these exclusion criteria depended on what the authors had available in their data, this was an ad-hoc procedure.
Table 5.1: Included Test Variables.

<table>
<thead>
<tr>
<th>Test variable</th>
<th>Abbreviation</th>
<th>Additional information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trail Making Test Part A</td>
<td>TMTA</td>
<td>Combined with Color Trails Test Part 1, D-KEFS Trail Making Test condition 2.</td>
</tr>
<tr>
<td>Trail Making Test Part B</td>
<td>TMTB</td>
<td>Combined with Color Trails Test Part 2, D-KEFS Trail Making Test condition 4.</td>
</tr>
<tr>
<td>Logical Memory I</td>
<td>LMI</td>
<td>Combined across multiple WMS versions, combined with RBANS Story Immediate Memory.</td>
</tr>
<tr>
<td>Logical Memory II</td>
<td>LMII</td>
<td>Combined across multiple WMS versions, combined with RBANS Story Delayed Memory.</td>
</tr>
<tr>
<td>Letter Fluency</td>
<td>LF</td>
<td>Synonyms: Controlled Oral Word Association Test, Phonemic Verbal Fluency.</td>
</tr>
<tr>
<td>Semantic Fluency</td>
<td>SF</td>
<td>Synonyms: Categorical Verbal Fluency. Preferential inclusion of the &quot;Animals&quot; version if multiple were available.</td>
</tr>
<tr>
<td>Digit Span Forwards</td>
<td>DSF</td>
<td>Combined across multiple WAIS and WMS versions.</td>
</tr>
<tr>
<td>Digit Span Backwards</td>
<td>DSB</td>
<td>Combined across multiple WAIS and WMS versions.</td>
</tr>
<tr>
<td>Coding</td>
<td>COD</td>
<td>Combined across multiple WAIS versions. Synonym: Digit Symbol Substitution.</td>
</tr>
<tr>
<td>Boston Naming Test</td>
<td>BNT</td>
<td></td>
</tr>
</tbody>
</table>
5.3.1.5 Analysis

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

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

5.3.1.6 Candidate factor models

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

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

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

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

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

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

5.3.2 Results

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

<table>
<thead>
<tr>
<th></th>
<th>TMTA</th>
<th>TMTB</th>
<th>LMI</th>
<th>LMII</th>
<th>LF</th>
<th>SF</th>
<th>DSF</th>
<th>DSB</th>
<th>COD</th>
<th>BNT</th>
<th>VLT-TR</th>
<th>VLT-DR</th>
</tr>
</thead>
<tbody>
<tr>
<td>One factor</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>Strauss</td>
<td>D</td>
<td>D</td>
<td>C</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>E</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Lezak</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Gross</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Hoogland</td>
<td>B</td>
<td>B + D</td>
<td>C</td>
<td>C</td>
<td>A + D</td>
<td>A + D</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Larrabee</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>D</td>
<td>D</td>
<td>A</td>
<td>C</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Jewsbury 1</td>
<td>B</td>
<td>B</td>
<td>A + C</td>
<td>A + C</td>
<td>E</td>
<td>E</td>
<td>D</td>
<td>D</td>
<td>B</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Jewsbury 2</td>
<td>B</td>
<td>B + D</td>
<td>A + C</td>
<td>A + C</td>
<td>E</td>
<td>E</td>
<td>D</td>
<td>D</td>
<td>B</td>
<td>A</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>


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

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

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

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

Figure 5.1: PRISMA diagram.

<table>
<thead>
<tr>
<th></th>
<th>TMTA</th>
<th>TMTB</th>
<th>LMI</th>
<th>LMII</th>
<th>LF</th>
<th>SF</th>
<th>DSF</th>
<th>DSB</th>
<th>COD</th>
<th>BNT</th>
<th>VLT-TR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMTB</td>
<td>0.543</td>
<td>0.012</td>
<td>-0.171</td>
<td>0.015</td>
<td>0.009</td>
<td>0.168</td>
<td>0.009</td>
<td>0.159</td>
<td>-0.185</td>
<td>-0.140</td>
<td>0.018</td>
</tr>
<tr>
<td>LMI</td>
<td>-0.084</td>
<td>-0.176</td>
<td>0.864</td>
<td>0.198</td>
<td>0.208</td>
<td>0.457</td>
<td>0.009</td>
<td>0.222</td>
<td>-0.256</td>
<td>-0.283</td>
<td>-0.195</td>
</tr>
<tr>
<td>LMII</td>
<td>-0.091</td>
<td>-0.264</td>
<td>0.069</td>
<td>0.134</td>
<td>0.231</td>
<td>0.168</td>
<td>0.009</td>
<td>0.473</td>
<td>0.272</td>
<td>0.230</td>
<td>0.481</td>
</tr>
<tr>
<td>LF</td>
<td>-0.207</td>
<td>-0.264</td>
<td>0.069</td>
<td>0.134</td>
<td>0.231</td>
<td>0.168</td>
<td>0.009</td>
<td>0.473</td>
<td>0.272</td>
<td>0.230</td>
<td>0.481</td>
</tr>
<tr>
<td>SF</td>
<td>-0.222</td>
<td>-0.256</td>
<td>0.262</td>
<td>0.274</td>
<td>0.457</td>
<td>0.481</td>
<td>0.007</td>
<td>0.185</td>
<td>0.258</td>
<td>0.267</td>
<td>0.481</td>
</tr>
<tr>
<td>DSF</td>
<td>-0.117</td>
<td>-0.202</td>
<td>0.151</td>
<td>0.231</td>
<td>0.351</td>
<td>0.348</td>
<td>0.187</td>
<td>0.271</td>
<td>0.300</td>
<td>0.232</td>
<td>0.300</td>
</tr>
<tr>
<td>DSB</td>
<td>-0.159</td>
<td>-0.283</td>
<td>0.236</td>
<td>0.220</td>
<td>0.272</td>
<td>0.230</td>
<td>0.481</td>
<td>0.185</td>
<td>0.258</td>
<td>0.267</td>
<td>0.481</td>
</tr>
<tr>
<td>COD</td>
<td>-0.487</td>
<td>-0.516</td>
<td>0.239</td>
<td>0.256</td>
<td>0.351</td>
<td>0.348</td>
<td>0.187</td>
<td>0.271</td>
<td>0.300</td>
<td>0.232</td>
<td>0.300</td>
</tr>
<tr>
<td>BNT</td>
<td>-0.185</td>
<td>-0.195</td>
<td>0.258</td>
<td>0.267</td>
<td>0.234</td>
<td>0.284</td>
<td>0.128</td>
<td>0.153</td>
<td>0.304</td>
<td>0.241</td>
<td>0.241</td>
</tr>
<tr>
<td>VLT-TR</td>
<td>-0.154</td>
<td>-0.217</td>
<td>0.447</td>
<td>0.461</td>
<td>0.267</td>
<td>0.349</td>
<td>0.196</td>
<td>0.271</td>
<td>0.300</td>
<td>0.232</td>
<td>0.300</td>
</tr>
<tr>
<td>VLT-DR</td>
<td>-0.140</td>
<td>-0.154</td>
<td>0.439</td>
<td>0.502</td>
<td>0.197</td>
<td>0.322</td>
<td>0.092</td>
<td>0.183</td>
<td>0.278</td>
<td>0.225</td>
<td>0.695</td>
</tr>
</tbody>
</table>

The results of the model comparison between candidate models is given in Table 4. The Hoogland et al. (2017), Lezak et al. (2012), and Strauss et al. (2006) models did not converge. Therefore the fit measures for these models should be interpreted with caution. With respect to relative fit, the AIC and BIC indicate that the two variants of the complex Jewsbury model fit better than the other models. With respect to absolute fit, the fit measures generally agree about the ordering of the models as well. All \( \chi^2 \) values indicate bad fit (all \( \chi^2 / df > 3 \)), which suggests that none of the models provides exact fit. All RMSEA values indicate good fit (all RMSEA < 0.05), except for the one factor model, where the RMSEA value indicates acceptable fit (RMSEA < 0.08). The SRMR values indicate bad fit for the five simplest models (SRMR > 0.10), and acceptable fit for the models described by Larrabee and Jewsbury et al. (SRMR > 0.05). The CFI values indicate bad fit for the one factor model (CFI < 0.95), acceptable fit for the model used by Gross et al. (CFI < 0.97), and good fit for the other models (CFI > 0.97).

The best-fitting Jewsbury model is depicted in Figure 2, in which correlations between latent variables are also provided. Because Trail Making Test A and B are measured in time to completion, these variables and the "Processing Speed" factor that they loaded on, are reverse coded. Therefore, the negative correlations between "Processing Speed" and the other latent factors should be interpreted such that better "Processing Speed" is correlated with better scores on the other latent factors.
Figure 5.2: Jewsbury 2 model for the twelve tests included in study 1. For each combination of latent factors, the correlation is given.
5.3.3 Discussion

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

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

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

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

Third, there was some overlap in the studies that were used in Jewsbury et al. (2016) and the studies that were included in this factor meta-analysis, so the sample that was used to develop the model was not completely distinct from the sample used to evaluate its performance. Therefore, the two analyses were not independent, which could have artificially improved the performance of the CHC model.
To address these issues, in the next study, the factor models will be fitted to raw data from the Netherlands and Belgium, combined in the ANDI database. This database allows us to use a single test version for every variable, and to use a single standardized education scale. Also, because raw data are available, we can directly incorporate the influence of demographic variables on test variables, rather than using the more indirect approach of partialing out these variables from the correlations. Last, this is a completely different sample of studies from the samples used in study 1, and the samples used by Jewsbury et al. (2016).

5.4 STUDY 2: FACTOR ANALYSIS OF THE ANDI DATABASE

5.4.1 Methods

5.4.1.1 Sample

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

5.4.1.2 Tests

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

5.4.1.3 Model changes

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

<table>
<thead>
<tr>
<th></th>
<th>TMTA</th>
<th>TMTB</th>
<th>LMI</th>
<th>LMII</th>
<th>LF</th>
<th>SF</th>
<th>COD</th>
<th>BNT</th>
<th>VLT-TR</th>
<th>VLT-DR</th>
</tr>
</thead>
<tbody>
<tr>
<td>One factor</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>Strauss</td>
<td>D</td>
<td>D</td>
<td>C</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>E</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Lezak</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Gross</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Hoogland</td>
<td>B</td>
<td>B + D</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>D</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Jewsbury</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>C</td>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>


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

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

5.4.2 Results

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

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

The $\chi^2$ values indicates bad fit for all models ($\chi^2 / df > 3$), except for the Jewsbury model, for which fit was acceptable ($\chi^2 / df > 2$). All RMSEA values indicate good fit (all RMSEA < 0.05), except for the one factor model, for which the RMSEA indicates acceptable fit (RMSEA < 0.08). The SRMR values indicate bad fit for the one factor and Hoogland models (SRMR > 0.10), and acceptable fit for the
In this article, we sought to establish the cognitive domains that are measured by neuropsychological tests. Cognitive domains are used on a daily basis by neuropsychologists, to make decisions on which tests to administer to a particular patient, to determine whether a disorder affects a single domain or multiple domains, to calculate com-
92 cognitive domains in neuropsychology: support for the CHC model

Figure 5.3: Jewsbury model for the ten tests included in study 2. For each combination of latent factors, the correlation is given for the meta-analytic data in roman type, and for the ANDI data in italic type.
posite scores of different tests belonging to the same domain, and to validate new tests that are designed to measure a particular cognitive function.

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

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

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

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

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

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

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

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

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

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