Statistical advances in clinical neuropsychology
Agelink van Rentergem Zandvliet, J.A.

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: http://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.
MULTIVARIATE NORMATIVE COMPARISONS FOR NEUROPSYCHOLOGICAL ASSESSMENT BY A MULTILEVEL FACTOR STRUCTURE OR MULTIPLE IMPUTATION APPROACH

4.1 ABSTRACT

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

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

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

4.2 INTRODUCTION

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

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

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

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

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

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

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

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

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

4.2.1 Between study variance

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

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

4.2.2 Structurally missing data

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

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

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

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

4.2.3 Missing combinations of tests

Because the normative database is composed of studies that have already been conducted, there is no control over which tests are administered to whom. In some studies, memory will have been the primary focus, while other studies may focus on executive functions. Therefore, some tests will be administered together in many studies, and some tests will never be administered together. Chances are that such missing combinations of tests will always exist, even if a large number of studies are included.
A multivariate normative comparison uses the covariance between tests in determining whether a profile of scores is abnormal. If a combination of two tests has not been administered, calculating the covariance between these two tests is not straightforward. We here consider two potential solutions: using a multiple imputation approach, or a factor structure approach.

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

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

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

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

4.3.1 Multivariate normative comparisons

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

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

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

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

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

where $x_{1i}, x_{2i}$ and $x_{3i}$ are scores on background variables for patient $i$, which in neuropsychological settings often correspond to sex, age and level of education, $\otimes$ denotes the Kronecker-product, $I$ is an identity matrix of size $P \times P$, $\beta_0$ is a column vector of length $P$, containing the intercepts for every test. These intercepts can be interpreted as the test scores that are predicted if the values of sex, age and level...
of education are all equal to zero. $\beta_1$, $\beta_2$ and $\beta_3$ are column vectors of length $P$, which contain the effects of a background variable on each test. For example, $\beta_1$, gives the change in test score for each test that results from a one unit increase in background variable $x_1$.

4.3.2 Modeling

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

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

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

$$S = \text{COV}_{u0} + \text{COV}_\epsilon \quad (4.4)$$

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

$\text{COV}_\epsilon$ is a covariance matrix of size $PxP$ of within study residual elements $\epsilon_{i1}$ to $\epsilon_{ip}$. The elements on the diagonal of the covariance matrix represent the within study variances of the test scores. The off-diagonal elements represent the within study covariances between test scores. An example of the structure for the residual covariance matrix $S$, if we leave $\text{COV}_\epsilon$ completely unstructured, is given in Table 1.
An advantage of this unstructured approach is that every within study covariance is estimated freely, which allows for any pattern of correlations that may exist in the data. However, this model is not identified when combinations of tests are missing. If tests 1 and 4 are not administered to the same participants, no single value can be identified for $\text{cov}(\epsilon_4, \epsilon_1)$. However, it might be argued that we may still come to an estimate of $\text{cov}(\epsilon_4, \epsilon_1)$, if we use Multiple Imputation (MI).

### 4.3.3 Multiple Imputation

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

### 4.3.4 Factor structure

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

$$\text{COV}_\epsilon = \Lambda \Psi \Lambda' + \Theta \quad (4.5)$$
where $\Lambda$ is a factor loading matrix of size $P \times M$ relating $P$ test scores to $M$ latent factors, and $\Lambda'$ is the transpose of this matrix. $\Psi$ is a matrix of size $M \times M$ containing the variances of the latent factors on the diagonal, and the covariances between latent factors on the off-diagonal. $\Theta$ is a diagonal matrix of size $P \times P$, containing the within study variances in test scores that are not explained by the latent factor or background variables. $\Theta$ is diagonal because given the latent factors, the tests should no longer be correlated (Bollen, 2002). If a test variable is included as an indicator for a latent factor, but is not correlated with the other test variables, the factor loading for this variable will be estimated to be 0. In this case, all variance in this test variable is not due to the latent factor, but is specific to this variable.

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

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

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

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

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

Note that each off-diagonal element of the matrix in Table 2 is estimable, even if a combination of tests has not been administered to the same participants (Cudeck, 2000). Therefore, an important advantage of the factor structure approach is that it allows for estimation of covariance between tests that have not been administered together. Another advantage is that this gives a parsimonious description of the data, which may enhance the sensitivity of normative comparisons. A disadvantage of the factor model approach is that it assumes correct model specification.
Table 4.2: Residual Covariance Matrix $S$ for Four Tests and One Latent Factor

<table>
<thead>
<tr>
<th>test</th>
<th>test 2</th>
<th>test 3</th>
<th>test 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>test 1</td>
<td>$\text{var}(u_{01}) + \lambda_{11}\psi_{11}\lambda_{11} + \theta_{11}$</td>
<td>$\lambda_{11}\psi_{11}\lambda_{21}$</td>
<td>$\lambda_{11}\psi_{11}\lambda_{31}$</td>
</tr>
<tr>
<td>test 2</td>
<td>$\lambda_{21}\psi_{11}\lambda_{11}$</td>
<td>$\text{var}(u_{02}) + \lambda_{21}\psi_{11}\lambda_{21} + \theta_{22}$</td>
<td>$\lambda_{21}\psi_{11}\lambda_{31}$</td>
</tr>
<tr>
<td>test 3</td>
<td>$\lambda_{31}\psi_{11}\lambda_{11}$</td>
<td>$\lambda_{31}\psi_{11}\lambda_{21}$</td>
<td>$\text{var}(u_{03}) + \lambda_{31}\psi_{11}\lambda_{31} + \theta_{33}$</td>
</tr>
<tr>
<td>test 4</td>
<td>$\lambda_{41}\psi_{11}\lambda_{11}$</td>
<td>$\lambda_{41}\psi_{11}\lambda_{21}$</td>
<td>$\lambda_{41}\psi_{11}\lambda_{31}$</td>
</tr>
</tbody>
</table>

4.3.5 Using the model estimates in comparisons

After the model is fitted, either using multiple imputation or using a factor structure specification, the total covariance matrix of tests can be calculated. With this covariance matrix, normative comparisons can be made for a particular patient. Generally, the patient will have completed fewer tests than the normative sample, and only some elements of the estimated vectors and matrices will be relevant for the normative comparison. For the tests that the patient has completed, the relevant elements from the covariance matrix $S$ are selected, and the relevant elements of the vectors $\beta_0$, $\beta_1$, $\beta_2$ and $\beta_3$ are selected to calculate the patient’s predicted test scores $\hat{y}_i$ from eq. 1.

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

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

4.4 SIMULATION STUDY

4.4.1 Outcomes

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

4.4.2 Parameter settings

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

4.4.3 Simulation conditions

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

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

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

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

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

### 4.4.4 Fitted factor model specification

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

### 4.4.5 Multiple Imputation settings

In the MICE procedure, a number of choices had to be made. First, a multilevel regression model with normally distributed errors was chosen to impute values, with within study variances assumed to be equal over studies. Multilevel models that allow within study variances to differ between studies show better imputation results (van Buuren, 2011), but require observations for every study for every test, which are not available for the current application. Second, background variables, and tests that were moderately or more highly correlated with the test to be imputed (i.e. $r \geq .10$, the default in the *mice* package, van Buuren, & Groothuis-Oudshoorn, 2011), were included as predictors in the imputation models. Third and fourth, the imputation algorithm ran for 10 iterations, and 50 complete datasets were
The results in the *false positive* condition are presented in Figure 2. If *one factor* was simulated, irrespective of the pattern of missing data, the proportion of significant results was always close to the required .05 level. If *two factors* were simulated, irrespective of the pattern of missing data, the false positive rate with a *two factor* model or an unstructured model was always close to nominal. However, if the *one factor* model was fitted to *two factor* data, the false positive rate became unacceptably large, for all missing data patterns.

The results in the *sensitivity* condition are presented in Figure 3. In the sensitivity condition, if *one factor* was simulated, the *one factor* model outperformed the unstructured model. The advantage of the *one factor* model increased with the degree of missing information, as shown in Figure 3. The same holds true if *two factors* were simulated and modeled. Note however that the high sensitivity when fitting a *one factor* model to *two factor* data should be interpreted in the light of the elevated false positive rate and can therefore not be celebrated.
We examined covariance parameters to investigate why sensitivity was higher for the one factor and two factor conditions than for the unstructured condition, and why the false positive rate was increased when fitting the one factor model to two factor data. Specifically, the covariance between tests 1 and 4 is of interest, as this is the parameter that should become more difficult to estimate when the extent of missing combinations worsens. These covariance estimates are plotted for every condition in Figure 4.

As shown in Figure 4, fitting the one factor model to two factor data leads to overestimates of the covariance, regardless of the pattern of missing data. This explains the increase in false positive rate observed in this condition. Fitting the one factor model to one factor data, and the two factor model to two factor data, led to correct estimation of covariances, regardless of the missing data pattern. For the unstructured condition, this was not the case, as the extent of the covariance underestimation increased when the combinations of test scores decreased. In the unstructured-disjunct condition, this covariance was even estimated as 0. This explains the drop in sensitivity that was observed between factor and unstructured conditions.
Figure 4.4: Density plot of covariance estimates by simulated factor model, fitted model and missing data pattern. The black dashed line indicates the covariance that was simulated. Note that in the unstructured disjunct condition all estimates were nearly zero, resulting in a very peaked distribution.
4.5.1 Additional simulations

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

4.5.2 Empirical example

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

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

In Mplus, the factor model was fitted, and the estimates of $\Lambda$, $\Psi$, $\Theta$, $COV_{g0}$ and $\beta$ were saved from the Mplus output. The lowest
Figure 4.5: Example five factor model. Observed variables, i.e. test variables and demographic variables in rectangles. Unobserved variables, i.e. latent factors and error terms, in circles. Single-headed arrows, black and gray, denote effects and factor loadings. Double-headed arrows denote correlations. $u$ and $\epsilon$ error terms are put in a single circle to simplify the representation, although the variance components are estimated separately.
N for these 27 variables was 153, so the numerator degrees of freedom would equal 27, and the denominator degrees of freedom would equal 126.

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

The multivariate normative comparison works as follows. First, the demographic variables and raw scores on the test variables are entered by the clinician. Second, the patient’s raw test scores are transformed and standardized to be on the same scale as the transformed and standardized norm data (de Vent et al., 2016). Second, predicted values are computed for all test variables, using equation 2 with the demographic values of the patient, and the regression coefficients $\beta$. Third, the covariance matrix of these test variables is computed, using equation 6 with the estimates $\Lambda$, $\Psi$, $\Theta$, $COV_{0g}$. Fourth, the MNC statistic is computed, using equation 1, with the patient data, the predicted values, and the covariance matrix. The output of the analysis is given in Table 4.

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

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

4.6 DISCUSSION

Multivariate normative comparisons provide a new neuropsychological tool that is sensitive to subtle deviations in a patient’s cognitive profile. However, these comparisons require that normative data are available for multiple tests from the same participants. We suggest to obtain the required normative data by providing a second life to data from control groups of neuropsychological studies, and aim to solve
Table 4.3: Simulated Scores on 27 Neuropsychological Test Variables.

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

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

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Sum of differences</th>
<th>Multivariate statistic</th>
<th>Degrees of freedom</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.46</td>
<td>1.77</td>
<td>27, 126</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Figure 4.6: Line plot of the standardized difference between the simulated patient data in Table 3 and the values predicted for this patient.
Figure 4.7: Example bivariate plot of the simulated patient data in Table 3, in comparison to the modeled normative data. The ellipse denotes 95% of the bivariate distribution. The triangle denotes the patient’s combination of scores. The dot denotes the predicted score for a participant of the patient’s sex, age and level of education.
several issues that arise when using such an aggregated neuropsychological database.

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

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

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

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

Third, it is assumed that the within study variances and covariances are equal over different studies, as there is no reason for variances of tests or covariances between tests to differ between studies.
However, in applications in which within study variances and covariances would differ between studies, this assumption can be relaxed.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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