Establishing normative data for multi-trial memory tests: the multivariate regression-based approach

Peer-reviewed author version

VAN DER ELST, Wim; MOLENBERGHS, Geert; van Tetering, Marleen & Jolles, Jelle (2017) Establishing normative data for multi-trial memory tests: the multivariate regression-based approach. In: CLINICAL NEUROPSYCHOLOGIST, 31(6-7), p. 1173-1187.

DOI: 10.1080/13854046.2017.1294202
Handle: http://hdl.handle.net/1942/26394

Establishing normative data for repeated cognitive assessment: a comparison of different statistical methods

Wim Van der Elst

LEARN! Research Institute and Faculty of Psychology & Education,

VU Universiteit Amsterdam (The Netherlands)

Geert Molenberghs

I-BioStat, Universiteit Hasselt and Katholieke Universiteit Leuven (Belgium)

Martin P. J. Van Boxtel

School for Mental Health and Neuroscience (MHeNS), and European Graduate School of Neuroscience (EURON), Maastricht University (the Netherlands)

Jelle Jolles

LEARN! Research Institute and Faculty of Psychology & Education,

VU Universiteit Amsterdam (The Netherlands)

Corresponding author: Wim Van der Elst, Faculty of Psychology & Education, VU Universiteit Amsterdam, Van der Boechorstraat 1, 1081 BT Amsterdam, The Netherlands (e-mail Wim.vanderelst@gmail.com).

Abstract

Serial cognitive assessment is conducted to monitor changes in the cognitive abilities of patients over time (e.g., to detect dementia). A problem with serial cognitive assessment is that the test scores at retesting occasions tend to increase due to practice effects. This should be taken into account when normative data are established.

At present, mainly the regression-based change and the ANCOVA approaches are used to establish normative data for serial cognitive assessment. However, these methods have some severe drawbacks. For example, they can only consider the data of two measurement occasions, and they cannot handle missing data appropriately.

In this paper, we propose three alternative methods that are not hampered by these problems (i.e., multivariate regression, the standard linear mixed model, and the linear mixed model combined with multiple imputation). The different methods are illustrated based on the data of a large longitudinal study in which the Stroop Color Word Test was administered at four subsequent measurement occasions. Based on these analyses and theoretical considerations, we recommend the use of the linear mixed model with multiple imputation, because (i) it allows for adequate modelling of the covariance structure (in contrast to the multivariate regression method), and because (ii) it takes the uncertainty of dealing with missing values into account (in contrast to the standard linear mixed model approach).

[GEERT] De standard linear mixed model approach, gebaseerd op direct likelihood, neemt die onzekerheid ook in overweging en in die zin is MI en direct likelihood (i.e., the standard approach) equivalent.

Word count abstract: 220 (max. 250)

Keywords: serial testing, norms, practice effects, longitudinal data, linear mixed model, multiple imputation, Stroop Color Word Test

Cognition is an umbrella term that refers to various higher-order behavioural abilities, such as memory, attention, and executive functions (Lezak, Howieson, & Loring, 2004). These higher-order behavioural abilities are latent variables that cannot be directly observed. Instead, they have to be inferred from proxy measures (Mitrushina, Boone, Razani, & D'Elia, 2005). For example, a person's verbal memory cannot be directly observed; what can be observed is the person's ability to recall verbal material that is presented in a specific standardized test setting.

Cognitive assessment is widely used in medical settings and in the behavioural sciences, for example in the context of diagnosing dementia (Lezak et al., 2004; Pasquier, 1999). In diagnostic settings, the "raw" score of a person on a cognitive test (e.g., the number of items that were recalled in a memory test) is usually not of direct interest. The reason for this is that the raw scores on cognitive tests are strongly affected by demographic variables (such as age and educational level; Mitrushina et al., 2005; Strauss, Sherman, & Spreen, 2006; Van der Elst, 2006). For example, the same raw test score may be indicative of a severe memory problem in a 50-year-old person, whilst it is within the normal limits of test performance for an 80-year-old person (Van der Elst, Van Breukelen, Van Boxtel, & Jolles, 2005). Clinicians therefore use relative measures (rather than raw test scores) to evaluate a patient's test performance (e.g., what is the percentage of demographically-matched "cognitively healthy" peers who obtain a test score that is equal to or worse than the test score of this patient?). So-called normative data are used to convert raw test scores into demographically-corrected relative measures (Mitrushina et al., 2005; Van der Elst, 2006).

In many diagnostic situations, the same cognitive test (or a parallel test version) is repeatedly administered to the same person. For example, a clinician may

need to determine whether a patient with early dementia has experienced cognitive decline since his or her last evaluation, or a clinician may need to evaluate whether a stroke patient has benefited from taking part in a rehabilitation program. Ideally, the observed changes in the test scores at subsequent measurement occasions would be directly interpretable in terms of true changes in the latent cognitive trait of interest. This is, however, generally *not* the case (Calamia, Markon, & Tranel, 2012). The main reason for this is that practice effects occur in serial testing situations. Practice effects refer to a variety of factors – such as procedural learning, memory for specific items, and increased comfort with formal testing situations (McCaffrey, Duff, & Westervelt, 2000) – that result in systematic improvements in test scores at retesting occasions, even though there was no true change in the latent trait that is measured by the cognitive test (Bartels, Wegrzyn, Wiedl, Ackermann, & Ehrenreich, 2010; Calamia et al., 2012; Dikmen, Heaton, Grant, & Temkin, 1999; Van der Elst, Van Breukelen, Van Boxtel, & Jolles, 2008; Temkin, Heaton, Grant, & Dikmen, 1999). Practice effects are especially pronounced when the test-retest intervals are short (e.g., Theisen, Rapport, Axelrod, & Brines, 1998), but they also occur when test-retest intervals of several years are used (Rönnlund & Nilsson, 2006; Salthouse, Schroeder, & Ferrer, 2004). In the latter case, the changes in the test scores over time reflect the combined influences of practice effects and true changes in the latent cognitive abilities (Van der Elst et al., 2008). Furthermore, the extent to which practice effects occur is affected by person characteristics such as the age and the educational level of a tested person (Mitrushina & Satz, 1991; Rapport, Brines, Axelrod, & Theisen, 1997; Stuss, Stethem, & Poirier, 1987; Van der Elst et al., 2008).

Failure to take practice effects into account may invalidate the conclusions that are drawn from a serial cognitive assessment (Calamia et al., 2012; Van der Elst

et al., 2008). For example, practice effects may mask the cognitive decline in a patient with early dementia, or practice effects may lead to the incorrect conclusion that a stroke patient has benefitted from a rehabilitation program. Normative data for serial cognitive assessment should thus take the testing history of a patient into account, but it is not clear which statistical method is optimal to achieve this aim (Heaton, Dikmen, Avitable, Taylor, Marcotte, & Grant, 2001; Temkin et al., 1999; Van der Elst et al., 2008).

Existing normative methods

In non-serial cognitive testing situations, normative data are established as based on classical univariate statistical methods. For example, an often-used procedure is the regression-based normative approach (Testa, Winicki, Pearlson, Gordon, & Schretlen, 2009; Van Breukelen & Vlaeyen, 2005; Van der Elst, Ouwehand, van Rijn, Lee, Van Boxtel, & Jolles, in press; Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2006a, 2006b, 2006c, 2006d). In this method, a classical multiple linear regression model is fitted to the data of a large sample of cognitively healthy people who were administered the cognitive test of interest (the normative sample). The multiple linear regression model assumes that $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{Y}_i$ is the vector of the outcomes, $\boldsymbol{X}_i$ is the design matrix (which typically includes age, gender, and educational level in normative studies), $\boldsymbol{\beta}$ is the vector of regression parameters, and $\boldsymbol{\varepsilon}_i$ is the vector of the residual components (for details on this model, see, e.g., Kutner, Nachtsheim, Neter, & Li, 2005).

Based on the established regression model, the test performance of a future patient *j* can be evaluated. This requires three steps. First, the expected test score of patient *j* is computed (i.e., $\hat{Y}_j = \boldsymbol{X}_j\widehat{\boldsymbol{\beta}}$). This score reflects the expected test score for a

cognitively healthy person who has the same demographic background as the tested patient. Second, the difference between the patient's observed and expected test scores is computed (i.e., $e_j = Y_j - \hat{Y}_j$) and standardized (i.e., $z_j = e_j / SD(e)$). The $SD(e)$ is the *SD* of the residuals in the normative sample (i.e., the positive square root of the residual mean squares). Third, the standardized residual of the patient is converted into a percentile value as based on the distribution of the standardized residuals in the normative sample. A percentile value below 5 is often considered as being indicative for a cognitive problem (because 95% of the "cognitively healthy" people perform better).

An important assumption of the classical linear regression model is that $\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\} = \sigma^2 \boldsymbol{I}$ (with $\boldsymbol{I}$ = an *n x n* identity matrix). Thus, it is assumed that the residuals (or equivalently, the outcomes) are uncorrelated. This assumption is not realistic in serial cognitive testing situations, because the cognitive test scores at subsequent measurement occasions tend to be highly correlated within individuals (Dikmen et al., 1999; Lezak et al., 2004; Temkin et al., 1999; Van der Elst et al., 2008). One possible solution to deal with this problem is to summarize the vector of the repeated measurements into change scores (change = endpoint score - baseline score), and subsequently regress these outcomes on the demographic covariates of interest in the normative sample (the regression-based change approach). Alternatively, the dependence issue can be solved by fitting a model in which the endpoint scores are regressed on the baseline scores and the demographic covariates in the normative sample (the ANCOVA approach).

Motivating example

The Maastricht Aging Study (MAAS) is a longitudinal research project into the determinants of cognitive aging (Jolles, Houx, Van Boxtel, & Ponds, 1995). The MAAS baseline measurement took place between 1993 and 1996, and three follow-up measurements were conducted (3, 6, and 12 years after baseline). All participants were thoroughly screened for medical pathology that could interfere with normal cognition (such as dementia or cerebrovascular disease).

The MAAS participants were administered an extensive battery of cognitive and medical tests. In the present paper, we will focus on the data of the Stroop Color Word Test (SCWT; Stroop, 1935). The SCWT is a well-known cognitive paradigm that is used to assess inhibition and other components of executive functioning (Lezak et al., 2004; Moering, Schinka, Mortimer, & Graves, 2004). The test consists of three subtasks. The first subtask shows colour words in random order (red, blue, yellow, green) that are printed in black ink. The second subtask displays solid colour patches in one of these four basic colours. The third subtask contains colour words that are printed in an incongruous ink colour (e.g., the word “red” printed in yellow ink). The participants were instructed to read the words, name the colours, and name the ink colour of the printed words as quickly and as accurately as possible in the three subsequent subtasks. The SCWT outcome variable of interest is the difference between the time that is needed to complete subtask three and the average time that is needed to complete the first two subtasks (i.e., SCWT score = time in seconds needed for subtask 3 - (time in seconds needed for subtasks 1 + 2) / 2). Higher SCWT scores are thus indicative for worse test performance.

In the MAAS, the SCWT was administered to N=887, N=696, N=614, and N=454 participants at the subsequent measurement occasions. Basic demographic data for the sample at baseline and at the three follow-up measurement occasions are

provided in Table 1. Level of Education (LE) was categorized into three levels using a classification scheme that is often used in the Netherlands (De Bie, 1987), with low = at most primary education, average = at most junior vocational training, and high = senior vocational or academic training. More details regarding the SCWT and the sample frame, participant recruitment, stratification criteria, and other aspects of the MAAS can be found elsewhere (Jolles et al., 1995; Van der Elst, 2006).

>>> Insert Table 1 about here <<<

Limitations of the existing normative methods

Suppose that the regression-based change method or the ANCOVA approach would be used to establish normative data for serial SCWT administration (as based on the MAAS data). This would have several major drawbacks.

First, the validity of the ANCOVA model depends on the assumption that there are no group differences in the baseline scores of the different demographic groups of interest (Kutner et al., 2005). This assumption is unrealistic in the context of cognitive assessment, because it has been consistently shown that age, gender, and educational level profoundly affect performance on the SCWT and on most other cognitive tests (Lezak et al., 2004; Mitrushina et al., 1999; Van der Elst, 2006).

Second, the ANCOVA and the regression-based change approaches cannot handle missing data appropriately. Both methods simply discard incomplete cases from the analyses, but a complete case analysis is only unbiased when the data are missing completely at random (MCAR; Little & Rubin, 1987; Verbeke & Molenberghs, 2000), and even then it is usually inefficient. MCAR means that the data are missing for reasons that are not related to the outcomes or to the

characteristics of the individuals. The MCAR assumption is not realistic in most serial testing settings. For example, the probability that a participant drops out of the MAAS is strongly affected by his or her baseline cognitive test score and age (Van Beijsterveldt, van Boxtel, Bosma, Houx, Buntinx, & Jolles, 2007), and consequently the MCAR assumption is not valid.

Third, the regression-based change and the ANCOVA methods can only use the data of a maximum of two measurement occasions. In the MAAS, the SCWT was administered four times. The application of the regression-based change or the ANCOVA approach would thus result in a substantial loss of information, and consequently a lowered precision of the parameter estimates and a loss of power (Verbeke & Molenberghs, 2000). Note that it might be argued that the endpoint score could be regressed on the test scores of multiple earlier testing occasions in the ANCOVA method (rather than on a single one), but this is generally not the case because the test scores at subsequent measurement occasions are highly correlated and thus collinearity issues would arise.

Alternative normative methods: the multivariate regression model, standard linear mixed model, and linear mixed model with multiple imputation

As noted in the previous section, the existing methods to establish normative data for serial cognitive assessment are fundamentally flawed. Applying these methods to the SCWT data (from the MAAS) would lead to a substantial loss of information and biased results. What we need are (i) methods that can deal with two or more correlated outcomes (within individuals), and (ii) methods that can handle missing data appropriately (without making unrealistic assumptions about the missingness mechanism).

Based on these criteria, the multivariate regression model, standard linear mixed model, and linear mixed model with multiple imputation are considered in the subsequent sections.

The multivariate regression model

The multivariate regression model assumes that $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$, with $\boldsymbol{Y}_i$: the vector of the repeated measurements, $\boldsymbol{X}_i$: the design matrix, $\boldsymbol{\beta}$: the vector of the regression parameters, and $\boldsymbol{\varepsilon}_i$: the vector of the error components. It is assumed that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, with $\mathbf{0}$: a zero matrix and $\boldsymbol{\Sigma}$: a general variance-covariance matrix of the residuals (for details on the model, see Johnson & Wichern, 2007).

In contrast to the classical, or univariate, linear regression model, the multivariate regression model can handle vectors of *repeated* observations for individuals. The parameter estimates in the multivariate regression model are based on likelihood methods, which allows for using all available data in the calculations. Moreover, the use of likelihood-based methods has the advantage that inferences can be based on the observed likelihood given a model that does not include a distribution for the missing data mechanism (so-called *ignorable* analyses; Little & Zhang, 2011; Molenberghs & Verbeke, 2005; Verbeke & Molenberghs, 2000). Note that ignorable analyses, when likelihood and Bayesian inferences are chosen, require that the missingness mechanism is missing at random (MAR, i.e., the missingness is independent of the unobserved data conditional on the observed data) or MCAR (as defined above), but this assumption can be relaxed in the context of normative analyses (see Discussion). Note also that the parameter estimates in a multivariate regression model can also be based on ordinary least squares methods (rather than on likelihood-based methods), but this situation will not be considered here because it

largely suffers from the same drawbacks as the regression-based change and the ANCOVA methods.

The standard linear mixed model

The standard Linear Mixed Model (LMM) assumes that $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$, with $\boldsymbol{Y}_i$: the vector of the repeated measurements, $\boldsymbol{X}_i$: the design matrix for the fixed effects (a.k.a. population-averaged parameters), $\boldsymbol{\beta}$: the vector of regression coefficients, $\boldsymbol{Z}_i$: the design matrix for the subject-specific effects (capturing how individuals deviate from the population average, where population is understood as anyone with the same fixed-effect design), $\boldsymbol{b}_i$: the vector of the random effects, and $\boldsymbol{\varepsilon}_i$: the vector of the residual components. It is assumed that $\boldsymbol{b}_i \sim N(\mathbf{0}, \mathbf{D})$ and that $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i)$, where $\mathbf{0}$ is a zero matrix and the $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_i$ are variance-covariance matrices (for details, see Verbeke & Molenberghs, 2000).

As compared to the multivariate regression model, the standard LMM has the additional advantage that both fixed and random effects can be included in the model. Random effects are not of substantive interest in normative studies (the focus is on the marginal evolutions, i.e., on the fixed effects), but it is nevertheless useful to model the covariance structure adequately because this generally leads to more efficient inferences for the fixed effects (i.e., smaller standard errors; Verbeke & Molenberghs, 2000). In addition, the LMM easily allows for unbalanced data, in the sense that it is possible for different subjects to provide different numbers of outcome values, either by design or because of missingness in the data.

The linear mixed model with multiple imputation

In the Linear Mixed Model with Multiple Imputation (LMM with MI) approach, the MI algorithm is first applied to fill in the missing observations in the dataset. The MI algorithm constructs multiple "completed" datasets as based on the incomplete dataset, by drawing from the conditional distribution of the unobserved outcomes given the observed ones (for details, see Beunckens, Molenberghs, & Kenward, 2005; Little & Rubin, 1987; Rubin, 1996; Molenberghs and Kenward 2007). Next, a LMM analysis is conducted on each of the completed datasets, and the different inferences are combined into a single one. As compared to the multivariate regression model and in line with the LMM, the LMM with MI has also the advantage that it can take the uncertainty of dealing with missing values into account in the analyses (Rubin, 1996; Verbeke & Molenberghs, 2000). This is to be contrasted with so-called single or simple imputation methods, where each missing value is substituted.

[GEERT] Ik heb het bovenstaande aangepast omdat er incorrect werd gesuggereerd dat het LMM die onzekerheid niet in overweging neemt, terwijl de beide methoden equivalent zijn in dit verband.

Application to the motivating example

In this section, we will illustrate the use of the multivariate regression, standard LMM, and LMM with MI methods to establish norms for serial SCWT administration (as based on the MAAS data). All analyses were conducted with R 2.14.0 for OS X and SAS v9.2 for Windows. An α-level of 0.05 was used.

The multivariate regression model

The initial multivariate regression model included the vector of the log(SCWT) scores as the outcome and age, age^2, gender, LE low, LE high, time and time^2 as the covariates. The SCWT score was log-transformed because preliminary analyses showed that the residuals were positively skewed. Age was centred (age = calendar age in years - 65) prior to the computation of the quadratic age effect (to avoid multicollinearity; Kutner et al., 2005). Gender was coded as 1 = male and 0 = female. The three levels of education (LE) were coded with two dummies, i.e., LE low: 1 = at most primary education and 0 = otherwise, and LE high: 1 = senior vocational or academic training and 0 = otherwise. Time was centred prior to the computation of the quadratic effects (time = time since baseline in years - 5.25). In addition to the main effects, the age x time, age x time^2, age^2 x time, age^2 x time^2, LE low x time, LE high x time, LE low x time^2, and LE high x time^2 interaction terms were included in the mean structure of the initial model. This was done because previous studies have suggested that older age and lower levels of education are associated with a more pronounced cognitive decline over time (see e.g., Salthouse, 1996; Schmand, Smit, Geerlings, & Lindeboom, 1997; Stern, 2003; Van der Elst et al., 2006a).

To obtain the most parsimonious model, it was evaluated whether the mean structure of the full model could be simplified by removing interactions and main effect terms. A series of classical likelihood ratio tests suggested that the model fit did not significantly deteriorate when the LE x time, age^2 x time, and time^2 covariates were removed from the model (see models 2 to 4 in Table 2). Next, the covariance structure was simplified by using a compound symmetry type (rather than an unstructured covariance type), but this resulted in a worse model fit (see model 5 in Table 2).

>>> Insert Table 2 about here <<<

The most parsimonious multivariate regression model that still adequately fitted the data was thus model 4. The parameter estimates for this model are provided in Table 3a. As shown, males and lower educated participants had significantly higher log(SCWT) scores at all measurement moments. There was a significant time x age interaction term, which suggested that the increase in the log(SCWT) scores over time was more pronounced for people who were older at baseline. The interaction is graphically depicted in Figure 1a for 50-, 65-, and 80-year-old females with an average LE (note that the shape of these plots is identical for males and for people with a low or a high educational level, i.e., the predicted log(SCWT) values are the same up to a constant). There was also a small (but significant) effect of age^2.

>>> Insert Table 3 about here <<<

>>> Insert Figure 1 about here <<<

The standard linear mixed model

The preliminary mean structure of the initial standard LMM was identical to the mean structure in the initial multivariate regression model (see above). A random intercept and two random slopes (for time and time^2) were included in the preliminary covariance structure (unstructured type). We first evaluated whether the random effects were all needed in the model, by removing one random effect after the other in a hierarchical way. Note that these tests cannot be conducted by using classical likelihood ratio procedures. Instead, a mixture of two χ^2 distributions should be used (with equal weights of 0.5; Verbeke & Molenberghs, 2000). The *p*-values of all the -

2l difference scores were significant (all $p < .05$; data not shown), indicating that the covariance structure could not be simplified by deleting random effects from the model.

Next, the non-significant fixed-effect terms were removed from the model (one after the other, in a hierarchical way) to obtain a more parsimonious mean structure. This procedure yielded a model which included age, age^2, gender, LE low, LE high, time, and the age x time interaction as the covariates (see models 2 to 7 in Table 4). It was subsequently evaluated whether the covariance structure could be simplified by using a compound symmetry structure, but this was not the case (see model 8 in Table 4).

>>> Insert Table 4 about here <<<

The parameter estimates for the final standard LMM (model 7) are presented in Table 3b. In agreement with the results of the multivariate regression model, being male and having a lower LE were associated with higher log(SCWT) scores at all measurement moments. There was again a significant age x time interaction, which suggested that the increase in the log(SCWT) scores over time was more pronounced for people who were older at baseline (see Figure 1b). The effect of age^2 was small but significant.

The linear mixed model with multiple imputation

The MI algorithm (Little & Rubin, 1987; Rubin, 1996) was used to replace each missing value by 10 different imputations. The final standard LMM (see Table 3b) was fitted in each of the 10 "complete" datasets, and the 10 inferences were combined

into a single one. There is an r statistic to quantify the uncertainty portion that is stemming from incompleteness, i.e., $r = \frac{(1+M^{-1})B}{\bar{u}}$ (where M = the number of imputations, B = the between-imputation variance, and $\bar{u}$ = the within-imputation variance; Schafer, 1999).

The final LMM with MI model is presented in Table 3c. The age x time and time parameters had the highest *r* values (i.e., 1.40 and 2.27, respectively). The *r* values for the other covariates were substantially lower and ranged between 0.07 and 0.63. The relatively large uncertainty for the age x time and the time covariates resulted in parameter estimates that were closer to zero and that had larger standard errors as compared to the results that were obtained for the multivariate regression model and the standard LMM (see Table 3).

[GEERT] Het bovenstaande is waarschijnlijk *geen* gevolg van de 'additional uncertainty' maar zou kunnen resulteren van imputation under the null. Als je age effecten wil bestuderen, dient age een variabele te zijn die mee is opgenomen in het imputatie-model, *zelfs als hij volledig is*. Zie ook de voorbeelden in onze 2005 en 2007 boeken. Anders gebeuren de imputaties "under the null" wat will zeggen dat je tijdens imputatie veronderstelt dat er geen verband is met age. Het is dan nadien niet verwonderlijk dat het effect van age (of age x time) schijnbaar daalt. Misschien moet daarom de multiple imputatie opnieuw gedaan worden.

As shown in Figure 1c, the LMM with MI model predicted a smaller increase in the log(SCWT) scores of older people over time, and a larger increase in the log(SCWT) scores of younger people over time. In agreement with the results of the multivariate regression model and the standard LMM, being male and having a lower

LE were associated with higher log(SCWT) scores at all measurement moments. The effect of age^2 was again small but significant. The parameter estimates and standard errors for the age, age^2, gender, and LE parameters in the LMM with MI model were similar to the values that were obtained for the multivariate regression model and the standard LMM.

Obtaining normative data

Analogously to the classical regression-based normative approach (see Introduction), three steps are needed to convert a future patient's log(SCWT) scores into percentile values. First, the expected log(SCWT) scores of patient *j* at time *t* are computed (=$\hat{Y}_{tj}$). Time *t* refers to the number of years since baseline. These calculations are based on the parameter estimates of the fixed effects that were provided in Table 3.

Second, the differences between the actually observed log(SCWT) scores of patient *j* at time *t* and the corresponding expected test scores are computed (i.e., $e_{tj} = -(Y_{tj} - \hat{Y}_{tj})$) and standardized (i.e., $z_{tj} = e_{tj}/SD(e_t)$). Note that the sign of the residuals is reversed here because a higher SCWT score is indicative of worse test performance. The $SD(e_t)$ values are the standard deviations of the residuals at time *t* in the normative sample. These values are presented in Table A1 (in Appendix).

Third, the standardized residuals (i.e., z_{tj}) are converted into percentile values. Histograms and QQ-plots suggested that the standardized residuals for the different models at all measurement moments were normally distributed in the MAAS (Figures not shown), and Kolmogorov-Smirnov tests supported this conclusion (all *p*-values > .098). The standardized residuals can thus be converted into percentile values by means of the standard normal distribution.

An example. Suppose that a 75-year-old average educated woman who is at risk for developing frontotemporal dementia is monitored over time. The patient was administered the SCWT at a baseline moment and 3, 6, and 12 years later. At the subsequent measurement occasions, she obtained SCWT test scores that equalled 80, 85, 90, and 100. The patient's $\log(SCWT_0)$, $\log(SCWT_3)$, $\log(SCWT_6)$, and $\log(SCWT_{12})$ scores thus equalled 4.382, 4.442, 4.500, and 4.605, respectively.

The clinician uses the LMM with MI approach to evaluate the patient's test performance. This requires three steps. First, the expected $\log(SCWT_0)$ test score is computed as based on Table 3c, i.e., 4.030 (= 3.876+(10*0.023)+((10^2)*0.0005)+(-5.25*0.017)+((10*-5.25)*0.0007)). Second, the standardized residual is computed (as based on Table A1 in Appendix), i.e., -1.067 (= -(4.382-4.030)/0.33). Third, the standardized residual is converted into a percentile value by means of the standard normal distribution. A standardized residual that equals -1.067 corresponds to a percentile value of 14. Thus, 14% of the population of 75-year-old cognitively healthy females with an average level of education obtain a $\log(SCWT_0)$ score that is equal to or higher than the score that was obtained by this woman. Using the same three-step procedure, the patient's $\log(SCWT_3)$, $\log(SCWT_6)$, and $\log(SCWT_{12})$ scores were normed. This yielded percentile values equal to 17, 20, and 26, respectively. Thus, the SCWT test performance of the patient is within normal limits at all the measurement moments.

User-friendly normative tables. A clinician can norm the test scores of a patient by performing the required computations by hand (as was illustrated in the previous paragraph), but this procedure is time consuming and prone to making errors. To increase the user-friendliness of the normative data for clinical use, we

established normative tables that present the raw SCWT scores that correspond to percentiles 5, 10, 25, 50, 75, 90, and 95, stratified by age (50, 55, ... , 80 years), gender, and LE (the normative tables can be downloaded at http://home.deds.nl/~wimvde/ [1]). The use of the normative tables is straightforward. For example, Table 1 in the online document immediately shows that the $SCWT_0$ score equal to 80 that was obtained by the 75-year-old average educated women of the previous example corresponds to a percentile value between 10 and 25. Note that the normative tables are based on the LMM with MI approach, because this method has some substantial advantages over the other methods (see Introduction and Discussion).

An automatic scoring program. The normative tables are easy-to-use but lack some accuracy, because (i) the tested patient's age has to be rounded-off if he or she is not exactly 50, 55, ... , 80 years old, and (ii) because only a limited number of percentile values can be presented in the normative tables (to limit their size to a convenient format). To maximize both the user-friendliness and the accuracy of the normative data, the normative conversion procedure was implemented into an Excel worksheet (which can be downloaded at http://home.deds.nl/~wimvde/ [2]). The use of the worksheet is straightforward: the clinician simply types in the age, gender, and LE of the tested patient together with his or her obtained raw SCWT scores at the different measurement moments, and the worksheet automatically computes the corresponding percentile values (based on the LMM with MI approach).

[1] Reviewer note: the normative tables will be placed on this website after publication of the present paper. For reviewing purposes, it can be downloaded from https://dl.dropbox.com/u/8416806/Serial%20Testing/Tables.pdf

[2] Reviewer note: the worksheet will be placed on this website after publication of the present paper. For reviewing purposes, it can be downloaded from https://dl.dropbox.com/u/8416806/Serial%20Testing/Norms.xls

Discussion

The multivariate regression model or the standard linear mixed model?

The standard LMM and the multivariate regression approaches yielded very similar results in the present study (see Tables 3a and b), but this will not always be the case. Especially when the data are highly imbalanced (e.g., when the repeated measurements are taken at widely varying time points), the results of both methods may differ more substantially (because the standard LMM approach allows for a more adequate modelling of the covariance structure than the multivariate regression method; Verbeke & Molenberghs, 2000). In the context of normative analyses, the use of an inappropriate covariance structure could lead to a situation where not all the relevant demographic covariates are taken into account in the construction of the normative data (not because the covariates are unimportant, but due to Type II error). The standard LMM approach is thus generally preferred over the multivariate regression method.

The standard linear mixed model or the linear mixed model with multiple imputation?

[GEERT] Gebaseerd op mijn eerdere commentaren is deze hele paragraaf fout. DL—LMM (direct likelihood LMM) en MI-LMM zijn equivalent (indien het multiple imputation model alle potentiele relaties bevat die later in het model worden opgenomen). Het is ook zo dat het standaard LMM een effect toelaat bij patiënten op momenten dat ze niet gemeten zijn (voor voorbeelden van hoe dat werkt, zie ook het boek van 2007). In situaties waar zowel responsen als covariaten ontbreken kan het MI gewoon handiger zijn om mee te werken, maar verder verschil is er niet. Het

voordeel van DL-LMM is dat het heel makkelijk uit te voeren is, en geen Monte Carlo component behoeft.

Is the missingness mechanism relevant when likelihood-based methods are used?

As noted in the Introduction, ignorable likelihood methods assume that the missingness mechanism is MCAR or MAR (as defined earlier). In the MAAS and in most other cognitive aging studies, the MCAR assumption is not valid (Van Beijsterveldt et al., 2002). Thus, the missingness mechanism is either MAR or MNAR (missing not at random, i.e., the missingness depends on unobserved data). A definitive test of MAR versus MNAR is not possible (because every MNAR model can be exactly reproduced by a MAR counterpart; Molenberghs, Beunckens, Sotto, & Kenward, 2008), but Verbeke, Molenberghs, and Rizopoulos (2010) argued that ignorable analyses provide reasonably stable results even when the MAR assumption is violated. The reason for this is that such analyses constrain the behavior of the unobserved data to be similar to the behavior of the observed data (Verbeke et al., 2010), and this is exactly what we want in the context of normative analyses. For example, suppose that a MAAS participant dropped out of the study at the second follow-up measurement occasion because he or she developed dementia. The missingness would clearly be associated with the unobserved $\log(SCWT_6)$ score (i.e., it would be MNAR), but this is not a problem because the unknown $\log(SCWT_6)$ score of the demented patient is not of interest. Indeed, in normative studies we are only interested in the test scores of cognitively healthy participants. When likelihood-based methods are used, the "unobserved" $\log(SCWT_6)$ and $\log(SCWT_{12})$ scores of the demented patient are modelled as based on the observed data of the patient at the

previous measurement moments (at which the patient was still cognitively healthy) and as based on the observed data at all measurement moments in the normative sample. As the observed data only include "cognitively healthy" individuals, appropriate estimates are obtained.

So, in the specific case of normative studies, the missingness mechanism is of less importance – at least when appropriate likelihood-based methods are used. As was noted in the Introduction, this is *not* the case when the regression-based change or the ANCOVA methods are used (i.e., the MCAR assumption is critical to obtain unbiased norms when these methods are used).

No Reliable Change Indices?

Early attempts to deal with practice effects and establish norms for serial testing situations consisted of computing so-called Reliable Change Indices with correction for practice (RCI; Chelune, Naugle, Lüders, Sedlak, & Awad, 1993, see also Jacobson & Truax, 1991). The RCI method uses the *overall* mean change score and the overall *SD*(change score) in a normative sample to establish confidence intervals for change scores. By comparing the change score of a patient with these upper and lower boundaries, it can be evaluated whether the patient's performance has changed significantly (i.e., declined or improved) over time.

We did not consider the RCI method in the present study, because it is merely a special case of the regression-based change method. Indeed, when the change score is not affected by any of the demographic covariates (in the normative sample), the final regression-based change model will only include the intercept (i.e., the overall mean change score), and the *SD*(residual) value will be equal to the overall *SD*(change score). Thus, apart from the general problems that hamper the validity of

the regression-based change method (see Introduction), the RCI method has the additional limitation that it makes the (unrealistic) assumption that the change scores are not affected by any of the demographic covariates.

Using non-linear models

The linear models that were used in the present study adequately described the evolution of the log(SCWT) scores over time. In most serial testing situations, linear models provide enough flexibility to capture the time trends of interest (e.g., quadratic or cubic time trends can be modelled by means of higher-order polynomials). Nonetheless, there are also a number of situations in which it might be useful to consider non-linear rather than linear models. For example, suppose that we would be interested in establishing normative data for the repeated quarterly measurement of phenomena that exhibit cyclic time trends (such as measures of immune functioning or depressive symptoms; Magnusson & Boivin, 2003; Nelson & Demas, 2004). Such phenomena cannot be easily modelled by means of linear models. Instead, a non-linear model might be preferred in which a sine (or cosine) function is used to capture the seasonal time trend. Apart from the fact that a non-linear model would be used to obtain the fixed effect estimates, the further normative procedure is identical to the three-step normative approach described earlier.

Correlated outcomes at a single time point

In the present paper, we focussed on the situation where the same cognitive test was repeatedly administered using large test-retest intervals, but the proposed methods are of course equally well applicable in situations where a number of correlated outcomes are available at the same measurement moment. For example, Rey's Verbal Learning

Test (VLT; Rey, 1958; Van der Elst et al., 2005) is a cognitive paradigm in which a sequence of fifteen words is presented to a participant in five subsequent learning trials. The increase in the number of recalled words over the five trials is often of clinical interest (for example, the learning curve is typically much more flat in demented patients than in cognitively healthy people), but it is unclear how the VLT data should be optimally analysed. At present, the vector of the five learning trial scores is typically summarized into a single measure (e.g., learning over trials = trial 5 score - trial 1 score), or separate analyses are conducted for each of the five trial scores (Lezak et al., 2004; Van der Elst et al., 2005). The summary measure approach is not optimal because a lot of the data is discarded. The separate trial score approach is also not optimal because it introduces multiple testing issues, and because the correlated nature of the data is not used in the analyses (whilst this information would be useful to model a person's VLT learning curve more adequately).

By using multivariate regression, standard LMM, or LMM with MI approaches, all five VLT trial scores can be included in a single parsimonious analysis which is not hampered by these problems. Similarly, norms can be obtained that are based on all available data (using the three-step normative procedure described above). Note that in this setting (i.e., when correlated outcomes are considered that were collected at a single time point), the differences in the results between the standard LMM and the LMM with MI approaches will be small because missing values are quite uncommon in non-serial testing situations. Similarly, the differences in the results between the multivariate regression and the LMM approaches will be small because the data are (almost) perfectly balanced.

General conclusion

At present, mainly the regression-based change and the ANCOVA approaches are used to establish normative data for serial cognitive assessment. These methods have the advantage that they are based on the classical linear regression model (which is well-known to most behavioural researchers and straightforward to perform), but they have some major disadvantages (e.g., they can only consider the data of two measurement occasions and they cannot deal with missing values in an appropriate way).

The multivariate regression, standard LMM, and LMM with MI approaches are not hampered by these problems.LMM and LMM-MI are largely equivalent, because they are valid under the same assumptions and neither artificially decrease nor increase the amount of information available. The advantage of the LMM is that it is easy to conduct and does not require a Monte Carlo component. LMM-MI on the other hand, flexibly deals with missing responses and missing covariates at the same time. It is important, when using MI, that all relationships (e.g., between covariates and responses) to be studied in the scientific model of interest, are present in the imputation model, to avoid "imputing under the null."

The log(SCWT) scores were affected by age, age^2, time, gender, and LE. These covariates should thus be taken into account in the construction of the normative data. There was also a significant time x age interaction, which suggested that the increase in the log(SCWT) scores over time was more pronounced for older people (as compared to their younger counterparts). These results are in line with previous findings in the cognitive aging literature (Salthouse, 1996; Schmand, Smit, Geerlings, & Lindeboom, 1997; Stern, 2003; Van der Elst et al., 2006a; Van der Elst, 2006). To increase the user-friendliness of the normative SCWT data, normative

tables and an automatic scoring program were provided (based on the results of the LMM with MI approach).

References

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience, 11,* 111-118.

Beunckens, C., Molenberghs, G., & Kenward, M. G. (2005). Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clinical Trials, 2,* 379-386.

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist, 26,* 543-570.

Chelune, G., Naugle, R. I., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology, 7,* 41-52.

De Bie, S. E. (1987). *Standaardvragen 1987: Voorstellen voor uniformering van vraagstellingen naar achtergrondkenmerken en interviews* [Standard questions 1987: Proposal for uniformization of questions regarding background variables and interviews]. Leiden, the Netherlands: Leiden University Press.

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test-retest reliability and practice effects of expanded Halstead-Reitan neuropsychological test battery. *Journal of the International Neuropsychological Society, 5,* 346-356.

Heaton, R. K., Temkin, N., Dikmen, S., Avitable, N., Taylor, M. J., Marcotte, T. D., & Grant, I. (2001). Detecting change: a comparison of three neuropsychological methods, using normal and clinical samples. *Archives of Clinical Neuropsychology, 16,* 75-91.

Jacobson, N. S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th. ed.). New York: Pearson Education, Inc.

Jolles, J., Houx, P. J., Van Boxtel, M. P. J., & Ponds, R. W. H. M. (1995). *Maastricht Aging Study: Determinants of Cognitive Aging.* Maastricht, the Netherlands: Neuropsych Publishers.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York: McGraw Hill.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological Assessment*. New York: Oxford University Press.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons.

Little, R. J., & Zhang, N. (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society Series C: Applied Statistics, 60,* 591-605.

Magnusson, A., & Boivin, D. (2003). Seasonal affective disorder: an overview. *Chronobiology International, 20,* 189-207.

McCaffrey, R. J., Duff, K., & Westervelt, H. J. (2000). *Practitioner's guide to evaluating change with neuropsychological assessment instruments.* New York: Kluwer Academic, Plenum Publishers.

Mitrushina, M. N., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York, NY: Oxford University Press.

Mitrushina, M., & Satz, P. (1991). Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology, 47,* 790-801.

Moering, R. G., Schinka, J. A., Mortimer, J. A., & Graves, A. B. (2004). Normative

data for elderly African Americans for the Stroop Color and Word Test. *Archives of Clinical Neuropsychology, 19,* 61-71.

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society, 70,* 371-388.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data.* Springer-Verlag: New York.

Nelson, R. J., & Demas, E. (2004). Seasonal patterns of stress, disease, and sickness responses. *Current Directions in Psychological Science, 13,* 198-201.

Pasquier, F. (1999). Early diagnosis of dementia: neuropsychology. *Journal of Neurology, 246,* 6-15.

Rapport, L. J., Brines, D. B., Axelrod, B. N., & Theisen, M. E. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *The Clinical Neuropsychologist, 11,* 375-380.

Rey, A. (1958). *L'examin clinique en psychologie.* Paris, France: Presses Universitaires de France.

Rönnlund, M., & Nilsson, L. G. (2006). Adult life-span patterns in WAIS-R Block Design performance: Cross-sectional versus longitudinal age gradients and relations

to demographic factors. *Intelligence, 34,* 63-78.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91,* 473-489.

Salthouse, T. A., Schroeder, D. H., & Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Developmental Psychology, 40,* 813-822.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103,* 403-428.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research, 8,* 3-15.

Schmand, B., Smit, J. H., Geerlings, M. I., & Lindeboom, J. (1997). The effects of intelligence and education on the development of dementia. A test of the brain reserve hypothesis. *Psychological Medicine, 27,* 1337-1344.

Stern, Y. (2003). The concept of cognitive reserve: A catalyst for research. *Journal of Clinical and Experimental Neuropsychology, 25,* 589-593.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91,* 473-489.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of neuropsychological tests: Administration, norms, and commentary* (3rd. ed.). New York: Oxford University Press.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643-662.

Stuss, D., Stethem, L., & Poirier, C. (1987). Comparison of three tests of attention and rapid information processing across six age groups. *The Clinical Neuropsychologist, 1,* 139-152.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society, 5,* 357-369.

Testa, S. M., Winicki, J. M., Pearlson, G. D., Gordon, B., & Schretlen, D. J. (2009). Accounting for estimated IQ in neuropsychological test performance with regression-based techniques. *Journal of the International Neuropsychological Society, 15,* 1012-1022.

Theisen, M. E., Rapport, L. J., Axelrod, B. N., & Brines, D. B. (1998). Effects of practice in repeated administrations of the Wechsler Memory Scale-Revised in normal adults. *Assessment, 5,* 85-92.

Van Beijsterveldt, C. E. M., Van Boxtel, M. P. J., Bosma, H., Houx, P. J., Buntix, F.,

& Jolles, J. (2007). Predictors of attrition in a longitudinal cognitive aging study: the Maastricht Aging Study (MAAS). *Journal of Clinical Epidemiology, 55,* 216-223.

Van Breukelen, G. J. P., & Vlaeyen, J. W. S. (2005). Norming clinical questionnaires with multiple regression: The Pain Cognition List. *Psychological Assessment, 17,* 336-344.

Van der Elst, W. (2006). *The Neuropsychometrics of Aging. Normative studies in the Maastricht Aging Study.* Maastricht, The Netherlands: Neuropsy publishers.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2005). Rey's Verbal Learning Test: Normative data for 1,855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society, 11,* 290-302.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006a). The Stroop Color-Word Test: influence of age, sex, and education; and normative data for a large sample across the adult age range. *Assessment, 13,* 62-79.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006b). The Concept Shifting Test: Adult normative data. *Psychological Assessment, 18,* 424-432.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006c). Normative data for the Animal, Profession and Letter M naming Verbal Fluency Tests

for Dutch speaking participants and the effects of age, education, and sex. *Journal of the International Neuropsychological Society, 12,* 80-89.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006d). The Letter Digit Substitution Test: Normative data for 1,858 healthy participants aged 24-81 from the Maastricht Aging Study (MAAS): influence of age, education, and sex. *Journal of Clinical and Experimental Neuropsychology, 28,* 998-1009.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2008a). Detecting the significance of changes in performance on the Stroop Color-Word Test, Verbal Learning Test of Rey, and Letter Digit Substitution Test after a test-retest interval of three years: the regression-based change approach. *Journal of the International Neuropsychological Society, 14*, 71-80.

Van der Elst, W., Ouwehand, C., van Rijn, P., Lee, N., Van Boxtel, M. P. J., & Jolles, J. (in press). The shortened Raven Standard Progressive Matrices: Item Response Theory-based psychometric analyses and normative data. *Assessment.*

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag: New York.

Verbeke, G., Molenberghs, G., & Rizopoulos, R. (2010). Random effects models for longitudinal data. In van Montfort, K., Oud, J., & Satorra, A. (Eds.). *Longitudinal Research with Latent Variables* (pp. 49-96). Springer-Verlag: New York.