Misuse of the Linear Mixed Model When Evaluating Risk Factors of Cognitive Decline

Cécile Proust-Lima*, Jean-François Dartigues, and Hélène Jacqmin-Gadda

* Correspondence to Dr. Cécile Proust-Lima, INSERM U897, Institut de Santé Publique, d’Épidémiologie et de Développement, Université Bordeaux Segalen, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France (e-mail: cecile.proust-lima@inserm.fr).

Initially submitted July 13, 2010; accepted for publication June 21, 2011.

The linear mixed model (LMM), which is routinely used to describe change in outcomes over time and its association with risk factors, assumes that a unit change in any predictor is associated with a constant change in the outcome. When it is used on psychometric tests, this assumption may not hold. Indeed, psychometric tests usually suffer from ceiling and/or floor effects and curvilinearity (i.e., varying sensitivity to change). The authors aimed to determine the consequences of such misspecification when evaluating predictors of cognitive decline. As an alternative to the LMM, they considered 2 mixed models based on latent processes that handle discrete and bounded outcomes. Model differences are illustrated here using data on 4 psychometric tests from the Personnes Âgées QUID (PAQUID) Study (1989–2004). The type I error of the Wald test for risk-factor regression parameters was formally assessed in a simulation study. It demonstrated that type I errors in the LMM could be dramatically inflated for some tests, such that spurious associations with risk factors were found. In particular, confusion between effects on mean level and effects on change over time was highlighted. The authors recommend use of the alternative mixed models when studying psychometric tests and more generally quantitative scales (quality of life, activities of daily living).

Abbreviations: AIC, Akaike’s Information Criterion; BVRT, Benton Visual Retention Test; CALC, calculation subscore of the MMSE; GA, general aging; IST, Isaacs Set Test; LMM, linear mixed model; MMSE, Mini-Mental State Examination; PAQUID, Personnes Âgées QUID; PDPD, prediagnostic phase of dementia.

In the study of chronic diseases, the linear mixed model (LMM) has become a standard method for describing the change over time in markers of progression and its association with risk factors. It offers a flexible statistical framework with which to dynamically describe disease progression through the trajectory of repeated marker data. The LMM relies on the assumptions that 1) the outcome of interest is continuous, 2) the random components of the model are Gaussian, and 3) a unit change in any predictor is associated with a constant fixed change in the outcome. It has been shown that inference with LMM is robust to violation of assumption 2 and especially misspecification of the random-effects distribution (1–3) or of the error distribution (4–6) when the mean structure is correct and assumption 3 holds.

In cognitive aging, the markers of progression are psychometric tests that are noisy measures of the underlying latent cognitive level, the biologic process of interest. They are usually discrete quantitative outcomes consisting of sum-scores of items and have specific properties. First, because of a limited range of possible values, they usually suffer from ceiling and/or floor effects (7). Moreover, in the range of possible values, they usually have a varying sensitivity to change (8) that we will refer to as curvilinearity. It means that a unit change on the psychometric test may not represent the same intensity of cognitive change in the underlying latent cognitive level scale at different levels of the psychometric test.

Despite these specific properties, which may result in markedly skewed and bounded test distributions and departures...
from the assumptions of the LMM, change in psychometric test scores over time is still usually studied through the standard LMM. Recent examples include evaluation of the effect of risk factors on change in the Mini-Mental State Examination (MMSE) (a widely used test measuring global cognitive performance) or other psychometric tests. Change over time of a pretransformation of the outcome was rarely analyzed, like the square root of the number of errors in the MMSE, the distribution of which was closer to the Gaussian distribution in a population of subjects free of dementia.

Alternatively, mixed models with latent processes can be used to account for the discrete nature and curvilinearity of the psychometric tests (including ceiling/floor effects) in longitudinal studies of cognitive aging. In these models, the latent process represents the actual unobserved cognitive level that underlies the psychometric test. Change over time of this latent process is described according to covariates in an LMM, and an equation of observation defines the link with the outcome. Using the item response theory, the latent process is directly related to the items constituting the sum-score with an item-specific equation of observation. However, in longitudinal settings, these models become very complicated and computationally intensive, with a large number of parameters, so their use has been limited until now. In contrast, the threshold model directly describes the sum-score by considering that each level of the sum-score corresponds to a specific interval of the latent process, the limits of the intervals being estimated. This approach, which corresponds to an item response theory model for a single graded item, takes into account the discrete and bounded nature of the psychometric test. However, its use has also been limited because it induces a large number of parameters and a numerical burden.

To avoid computational problems induced by item response models or threshold models, Proust et al. proposed estimating the nonlinear function linking the latent process level with the outcome through a parsimonious family of flexible continuous transformations. This approach extends the idea of pretransformation of the outcome to correct for curvilinearity and ceiling/floor effects by directly estimating the transformation that is the most adapted to the data along with the regression model. It was shown to markedly improve the goodness of fit in comparison with the LMM and to highlight metrologic properties of the psychometric tests while remaining computationally easy.

Despite these alternative models, which handle typical asymmetric and bounded psychometric test distributions, the LMM is still widely used without further checking, and results are interpreted without taking into account the limits of the analyses. Thus, our objective in this work was to evaluate the consequences of neglecting the potential curvilinearity of the tests when studying associations between risk factors and cognitive decline. Differences between the standard LMM and alternative latent process models are illustrated using data from the Personnes Agées QUID (PAQUID) cohort. In a simulation study, the type I errors of the Wald tests for the risk-factor regression parameters that are of main interest in such studies are assessed. Finally, recommendations for future epidemiologic studies of the impact of risk factors on cognitive decline are given.

**MATERIALS AND METHODS**

**Population**

The PAQUID Study is a French epidemiologic study relying on a population-based sample of 3,777 community-dwelling persons aged 65 years or older. Subjects were evaluated at home at an initial visit (V₀) in 1988–1989 and were followed up 7 times at years 1, 3, 5, 8, 10, 13, and 15 (hereafter called V₁–V₁₅). V₁₅ took place in 2003–2004. At each visit, a neuropsychological evaluation and a 2-phase screening procedure for diagnosis of dementia were carried out at home. A detailed description of the PAQUID program is presented by Letenneur et al. (23).

**Neuropsychological evaluation**

Four psychometric tests for which low values indicate more severe impairment were considered. These psychometric tests were chosen because they are largely used in epidemiology and illustrate different characteristics of psychometric tests: asymmetric distribution, ceiling/floor effect, and/or a small number of levels.

The MMSE evaluates various dimensions of cognition (memory, calculation, orientation in space and time, language, and word recognition). It is often used as an index of global cognitive performance, and the score ranges from 0 to 30. The calculation subscore of the MMSE (CALC) consists in subtracting iteratively 5 times the number 7, beginning from 100. The score ranges from 0 to 5. The recognition form of the Benton Visual Retention Test (BVRT) evaluates immediate visual memory. After a 10-second presentation of a stimulus card displaying geometric figures, subjects are asked to choose the initial figure among 4 possibilities. A total of 15 stimulus cards are successively presented, so the score ranges from 0 to 15. The Isaeas Set Test (IST) (25), shortened to 15 seconds, evaluates semantic verbal fluency and processing speed. Subjects are required to name words (with a maximum of 10) in 4 specific semantic categories (cities, fruits, animals, and colors) in 15 seconds. The score ranges from 0 to 40.

**Sample selection**

Cognitive measurements taken at V₀ were excluded from the analysis because of a learning effect between the first 2 examinations. Two participant samples were considered: persons in the prediagnostic phase of dementia (PDPD), in which only subjects with incident dementia between V₁ and V₁₅ were included and post-dementia-diagnosis data were excluded, and a heterogeneous sample representing general aging (GA), in which every subject who was free of dementia at V₁ was included. For any sample (PDPD or GA) and any psychometric test, subjects who had at least 1 measurement between V₁ and V₁₅ were included. This led to 6 samples comprising 2,897 subjects for MMSE and CALC in the GA sample (n = 612 in PDPD), 2,623 subjects for BVRT in the...
GA sample (n = 514 in PDPD), and 2,760 subjects for IST in the GA sample (n = 574 in PDPD).

Statistical models

All of the statistical models (including the standard LMM) are presented as latent process models, described in Figure 1. The latent process of interest, called Λ, represents the latent cognitive level at any time t that underlies the psychometric test. Change over time in Λ for subject i (i = 1,..., n) is described according to time t and covariate X; in a standard LMM (26):

\[ Λ_i(t) = β_0 + β_1 t + β_2 X_i + u_0 + u_1 t, \]

where \( u_0 \) and \( u_1 \) are the random intercept and slope that account for the correlation between repeated measures, respectively. They are correlated and follow a Gaussian distribution. At the population level, \( β_0 \) is the mean of the latent process for \( t = 0 \) and \( X = 0 \); \( β_1 \) is the mean slope representing the mean change in Λ in a time unit for \( X = 0 \); \( β_2 \) corresponds to the mean change in Λ at \( t = 0 \) for a unit change in \( X \); and \( β_3 \) corresponds to the mean change in the slope of Λ for a unit change in \( X \). In the following, \( β_0 \) and the variance of \( u_0 \) are respectively constrained to 0 and 1 for identifiability purposes.

For each statistical model, an equation of observation links the repeated measure of outcome \( Y_{ij} \) with the latent process level at the observation time \( t_{ij} \), with \( j \) representing the occasion (j = 1,..., \( n_i \)):

- The standard LMM is obtained by assuming \( Y_{ij} = a + b_0 Λ(t_{ij}) + ε_{ij} \), where \( ε_{ij} \) is an independent Gaussian measurement error at time \( t_{ij} \) and \( a \) and \( b \) are parameters needing to be estimated that replace \( β_0 \) and the variance of \( u_0 \).
- The standard LMM applied on a pretransformed outcome assumes similarly that \( h(Y_{ij}) = a + b_0 Λ(t_{ij}) + ε_{ij} \), where \( h(\cdot) \) is the pretransformation (e.g., \( h(\text{MMSE}) = \sqrt{30 - \text{MMSE}} \)) and \( ε_{ij}, a, \) and \( b \) are defined as above.
- In the latent process model proposed by Proust et al. (16) and called beta LMM, the equation of observation is \( h(Y_{ij}, η) = a + b Λ(t_{ij}) + ε_{ij} \), where the transformation \( h(\cdot, η) \) is a beta cumulative distribution function depending on parameters η that are estimated along with the regression parameters and \( ε_{ij}, a, \) and \( b \) are defined as above.
- In the threshold LMM, the equation of observation is defined at each level \( c \) of the outcome (\( c = 0, \ldots, C \)) by \( Y_{ij} = c \Leftrightarrow η_c \leq Λ(t_{ij}) + ε_{ij} < η_{c+1}, \) where \( ε_{ij} \) is defined as above, thresholds \( η_c (c = 1, \ldots, C) \) are estimated along with the regression parameters, and \( η_0 = -\infty \) and \( η_{C+1} = +\infty \) for identifiability (19).

Compared with the 3 former models, the threshold LMM is computationally more intensive because of the increased number of parameters and the numerical integration required in the log-likelihood computation. However, because it models directly each possible level of the outcome, it is considered the most adequate model among the candidates.

The goodnesses of fit of these 4 models were compared in the natural scale of the outcome using Akaike's Information Criterion (AIC) (27). To make possible the comparison of models assuming continuous (standard, pretransformed, and beta LMMs) and discrete (threshold LMM) data, the posterior likelihood was computed as the probability of observing the discrete values from estimated standard, pretransformed, and beta LMMs rather than the density (for details, see Appendix).

RESULTS

Illustration using data from the PAQUID cohort

Histograms of the 4 psychometric tests’ distributions are given in Figure 2 for the GA and PDPD samples. With the exception of the IST, they underline asymmetric and bounded distributions that could reflect curvilinearity problems and justify the use of more sophisticated mixed models.

Cognitive decline was studied according to age in the GA samples and time before dementia diagnosis in the PDPD samples, adjusted for educational level as a binary covariate (no diploma vs. graduation from primary school). The prevalence of subjects with no diploma varied from 29.9% to 32.3% in the GA samples and from 39.1% to 43.1% in the PDPD samples.

Figure 3 displays estimated transformations linking the psychometric test with the underlying Gaussian latent process in each model considered. Table 1 presents the corresponding AIC values.

For the MMSE, in both samples, the standard LMM assumed a linear transformation very far from the nonlinear one estimated by the most flexible threshold LMM. The difference in AIC (δAIC) between the models reached 4,227 points in the GA sample and 592 points in the PDPD sample. If the use of a pretransformation shrank this gap, it remained quite important, with a δAIC of 803 points in the GA sample and 77 points in the PDPD sample. In contrast, the estimated transformation from the beta LMM was close to the transformation stemming from the threshold LMM, showing that it constituted a good alternative. This was confirmed by largely reduced differences in AIC in the GA sample (delta AIC = 94) and the PDPD sample (delta AIC = 16).
Figure 2. Distributions of scores on 4 psychometric tests in the general aging sample (left) and the prediagnostic phase of dementia sample (right), Personnes Âgées QUID (PAQUID) Study, France, 1989–2004. Data were pooled from all available study visits. MMSE, Mini-Mental State Examination; BVRT, Benton Visual Retention Test; IST, Isaacs Set Test; CALC, calculation subscore of the MMSE. n, number of subjects. The median number of visits per subject for each test was 3 (interquartile range, 2–5).
Figure 3. Estimated transformations from latent process linear mixed models (LMMs) for the 4 psychometric tests according to the underlying Gaussian latent process in the general aging sample (left) and the prediagnostic phase of dementia sample (right), Personnes Agees QUID (PAQUID) Study, France, 1989–2004. Solid line, threshold LMM; long-dashed line, beta LMM; dotted line, standard LMM; short-dashed line, pretransformed LMM used for the Mini-Mental State Examination (MMSE) only. BVRT, Benton Visual Retention Test; IST, Isaacs Set Test; CALC, calculation subscore of the MMSE. Numbers of subjects and the median number of visits per subject are given in Figure 2.

For the BVRT, in the GA sample, the linear transformation assumed in the standard LMM was relatively far from the transformation estimated in the threshold LMM (with ΔAIC = 540). In the PDPD sample, transformations from the threshold and standard LMMs were relatively close in low levels of the tests. However, they differed in high levels, and the ΔAIC of 63 points indicated again a better fit of the threshold LMM. In contrast, the transformations from the beta LMM remained very close to the ones estimated in the threshold LMM in both samples, and the fits were very similar (the AIC of beta LMM was 2.4 points better in the PDPD sample and 6.4 points worse in the GA sample).

For the IST, the linear transformation assumed in the standard LMM was close to the ones estimated in the beta and threshold LMMs in both the GA and PDPD samples. In the GA sample, a slight difference in the transformations in very small IST values could be observed. Although this corresponded to few observations (see Figure 2), this led to a better AIC for the threshold model than for the beta (ΔAIC = 108) and standard (ΔAIC = 142) LMMs. In the PDPD sample, in contrast, the parsimonious beta and standard LMMs gave better AICs (ΔAIC = 19 and ΔAIC = 13, respectively), with a slight preference for the beta LMM.

For CALC, large differences in the estimated transformations from the standard and threshold LMMs were observed in the GA sample. They were reduced but still present in the PDPD sample. The ΔAIC reached 2,547 points in the GA sample and 434 points in the PDPD sample, which indicated the markedly better fit of the model, accounting for the discrete nature of the test. The beta LMM was not applied with CALC. Indeed, a model assuming continuous data is not appropriate for a test with 6 levels and may induce convergence problems (this also applies to the standard LMM).

These analyses showed that flexible models accounting for the test curvilinearity gave markedly better fits. Not surprisingly, these differences resulted in parameter estimates of the impact of educational level on baseline cognitive level and cognitive rate of change that were quite different (see Table 2). However, they also resulted in markedly different parameter $P$ values that even induced dissimilar conclusions regarding the association of educational level with cognitive decline (educational level $\times$ interaction). For example, in the GA sample, the threshold LMM accounting for curvilinearity did not highlight any interaction between educational level and cognitive decline for MMSE, BVRT, and CALC, while significant associations were found when using a standard LMM. These contradictions regarding the effect of educational level did not necessarily emerge in the predicted trajectories of cognitive decline represented in Figure 4: The trajectories remained relatively close in the threshold LMM and the standard LMM. Indeed, the curvilinearity means that the intensity of change represented by a loss of 1 point on the test varies as the test level changes. Consequently, although the distance in latent cognitive level between high and low educational levels remained the same across time and the range of latent cognition (no interaction with time), the distance between high and low educational levels in the test scale varied with time and the initial level of the test. When assuming a constant intensity of change in the entire range of the test’s scores—that is, using the standard LMM—this nonconstant gap between high and low educational levels was translated into a false interaction with time.

This kind of departure from the LMM assumptions is very difficult to highlight in diagnostic analyses. For example, except for MMSE and CALC in the GA sample, the subject-specific residuals from the standard LMM (displayed in Figure 5 with quantile-quantile plots) did not show any severe departure from the normality assumption. Results were the same for other diagnostic tests. Only the analysis of a better-suited model revealed the problem with curvilinearity.

**Simulation study**

Motivated by these observations, we performed a simulation study to evaluate the impact of misspecification of the
Two binary covariates were considered. MMSE, mim-BVRT, mim-IST, and mim-CALC, respectively. metric test distributions were mimicked and called mim-

The time scale was age for GA (transformed to 30 and retaining the same direction as MMSE), the beta LMM, and the threshold LMM. For mim-BVRT and mim-IST, only the standard, beta, and threshold LMMs were considered, since there were no usual pretransformations for these tests. For mim-CALC, only the standard and threshold LMMs were considered.

For each scenario, 500 samples of 500 subjects were simulated. Table 3 gives the estimated type I error (α) of the Wald test for the covariate effect on baseline (X) and interaction with time (X × t) for a nominal value of α = 5%. The estimated type I error of the Wald test was obtained by simulating the samples with a 0 parameter value and calculating the percentage of times where this parameter was found to be significantly different from 0 using the Wald test. This corresponds to testing a baseline effect in a model where there is only an interaction effect and inversely. In a well-specified model, α is close to 5%.

For mim-MMSE, α reached 93.2% in the GA sample and 97.8% in the PDPD sample for an interaction between X1 and time when we used the standard LMM. This indicated that the standard LMM concludes that there is an interaction 93.2% (or 97.8%) of the times when there is no interaction. When we used the pretransformation, α was reduced but was still far from the nominal value, with more than 40.8% for X1 × t and 13.0% for X2 × t. In contrast, using the beta LMM, α values were close to 5%. In any mixed model, α values for the X1 effect on baseline level were close to 5%. Indeed, because the interaction X1 × t was nearly 0, it could not blur the association with the baseline effect.

**Table 2. Estimates of the Association of Educational Level With Baseline Psychometric Test Score and Score Slope (Educational Level × Time (t)) for Different Mixed Models in 2 Participant Samples, PAQUID Study, France, 1989–2004**

<table>
<thead>
<tr>
<th>Psychometric Test and Estimated Model</th>
<th>General Aging Sample</th>
<th>Prediagnostic Phase of Dementia Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EL&lt;sup&gt;b&lt;/sup&gt;</td>
<td>P Value&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>Standard LMM</td>
<td>0.618</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Pretransformed LMM</td>
<td>0.909</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Beta LMM</td>
<td>1.026</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Threshold LMM</td>
<td>1.000</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Benton Visual Retention Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard LMM</td>
<td>1.067</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Beta LMM</td>
<td>1.033</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Threshold LMM</td>
<td>1.04</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Isaacs Set Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard LMM</td>
<td>1.031</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Beta LMM</td>
<td>1.038</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Threshold LMM</td>
<td>1.017</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>CALC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard LMM</td>
<td>1.241</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Threshold LMM</td>
<td>1.09</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Abbreviations: CALC, calculation subscore of the Mini-Mental State Examination; EL, educational level; LMM, linear mixed model; PAQUID, Personnes Âgées QUID.

<sup>a</sup> Numbers of subjects and the median number of visits per subject are given in Figure 2.

<sup>b</sup> Reference category, no diploma.

<sup>c</sup> Time t corresponds to age in decades from age 65 years in the general aging sample and years before diagnosis in the prediagnostic phase of dementia sample.

<sup>d</sup> P value from the Wald test.
Figure 4. Predicted trajectories of cognitive decline according to educational level for the 4 psychometric tests in the general aging sample (left) and the prediagnostic phase of dementia sample (right), derived using the standard linear mixed model (dashed line) and the threshold linear mixed model (solid line), Personnes Âgées QUID (PAQUID) Study, France, 1989–2004. In each graph, the top trajectories correspond to a higher educational level and the bottom trajectories correspond to a lower educational level. MMSE, Mini-Mental State Examination; BVRT, Benton Visual Retention Test; IST, Isaacs Set Test; CALC, calculation subscore of the MMSE. Numbers of subjects and the median number of visits per subject are given in Figure 2. Dotted lines, 95% confidence interval.
Figure 5. Quantile-quantile plots of subject-specific standardized residuals for the 4 psychometric tests in the general aging sample (left) and the prediagnostic phase of dementia sample (right) from the standard linear mixed model, Personnes Âgées QUID (PAQUID) Study, France, 1989–2004. MMSE, Mini-Mental State Examination; BVRT, Benton Visual Retention Test; IST, Isaacs Set Test; CALC, calculation sub-score of the MMSE. Numbers of subjects and the median number of visits per subject are given in Figure 2. Dashed lines, 95% confidence interval.
Table 3. Estimated Type I Error From the Wald Test (\(\hat{\alpha}\), Given in %) of a Regression Parameter for Covariate \(X_1\) or \(X_2\) and Interaction With Time (t) From 500 Replicated Samples of 500 Subjects in 2 Participant Samples, PAQUID Study, France, 1989–2004

<table>
<thead>
<tr>
<th>Outcome Distribution and Estimated Model</th>
<th>General Aging Sample</th>
<th>Prediagnostic Phase of Dementia Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(X_1)</td>
<td>(X_1 \times t^a)</td>
</tr>
<tr>
<td>mim-MMSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard LMM</td>
<td>5.0</td>
<td>93.2</td>
</tr>
<tr>
<td>Pretransformed LMM</td>
<td>3.6</td>
<td>40.8</td>
</tr>
<tr>
<td>Beta LMM</td>
<td>4.4</td>
<td>4.6</td>
</tr>
<tr>
<td>mim-BVRT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard LMM</td>
<td>3.2</td>
<td>21.4</td>
</tr>
<tr>
<td>Beta LMM</td>
<td>3.8</td>
<td>6.0</td>
</tr>
<tr>
<td>mim-IST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard LMM</td>
<td>4.6</td>
<td>8.0</td>
</tr>
<tr>
<td>Beta LMM</td>
<td>4.6</td>
<td>5.0</td>
</tr>
<tr>
<td>mim-CALC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard LMM</td>
<td>5.0</td>
<td>45.6</td>
</tr>
<tr>
<td>Threshold LMM</td>
<td>6.4</td>
<td>5.2</td>
</tr>
</tbody>
</table>

Abbreviations: LMM, linear mixed model; mim-BVRT, distribution mimicking the Benton Visual Retention Test; mim-CALC, distribution mimicking the calculation subscore of the MMSE; mim-IST, distribution mimicking the Isaacs Set Test; mim-MMSE, distribution mimicking the MMSE; MMSE, Mini-Mental State Examination; PAQUID, Personnes Âgées QUID.

* Time t corresponds to age in decades from age 65 years in the general aging sample and years before diagnosis in the prediagnostic phase of dementia sample.

\(\hat{\alpha}\) Estimated type I error (\(\hat{\alpha}\)) outside of the expected 95% interval 3.09, 6.91 around the nominal value of 5%.

The standard LMM worked better for mim-BVRT than for mim-MMSE. However, \(\hat{\alpha}\) reached 21.4% for some interactions between the covariate and the time. In contrast, the beta LMM again gave good results.

For mim-CALC, \(\hat{\alpha}\) values were high in the standard LMM for the interaction between \(X_1\) and time (45.6% in the GA sample and 56.4% in the PDPD sample) and for the \(X_2\) effect on baseline in the PDPD sample (25.0%), while the threshold LMM gave good results (it was the correct model).

Finally, for mim-IST, which highlighted a relatively linear transformation in Figure 3, \(\hat{\alpha}\) values were close to 5% in the standard and beta LMMs.

**DISCUSSION**

In this work, we aimed to demonstrate that the use of standard LMM on psychometric test data could lead to spurious associations of risk factors with cognitive change over time. Our conclusions were based on 4 psychometric tests that are largely used in epidemiologic studies and exhibit different types of distribution, as well as 2 distinct populations. With the exception of the IST, the illustration highlighted that the standard LMM provided a markedly worse fit of the data than the threshold LMM specifically adapted to discrete and bounded data or the beta LMM that accounted for curvilinearity while considering the data as continuous (16). Indeed, the estimated transformations derived from these models showed a clear nonlinear relation between each test and the underlying biologic process of interest, while the standard LMM assumed a linear transformation. This had been previously stated (8). The simulation study provided further arguments to demonstrate that for 3 of the psychometric tests, the associations with covariates in the standard LMM were distorted. In particular, when the risk factor had an effect on the initial cognitive level, the test for interaction with time was biased: It tended to find too often a spurious association between the risk factor and the rate of cognitive decline, while there was only an association with the initial level. This comes from the curvilinearity of the tests. The tests’ varying sensitivity to change induces a varying distance between 2 groups even when there is no interaction with time. By neglecting the test curvilinearity, the standard LMM interprets this varying distance as an interaction with time, so that confusion appears between the risk factor’s impact on the initial level and its impact on cognitive change with time. Such confusion can be of great importance. For example, if educational level, a proxy measure of the brain’s reserve capacity, were found to be related to decline on a given test, the investigators might conclude that reserve capacities were particularly linked to the underlying cognitive function without considering the misuse of the model. Curvilinearity constitutes an intrinsic property of the test that should be accounted for in any regression model whatever the studied population, covariates, or time scale. Especially, it cannot be corrected by changing the time scale or adding nonlinear covariate effects. For instance, the estimated nonlinear transformations in both PAQUID data sets were almost identical when we rescaled the latent process and were practically unchanged when we considered quadratic trajectories or removed the covariate (results not shown).

Recommendations should be addressed based on these findings. In any analysis evaluating the effect of risk factors on change in a psychometric test score over time, the standard
LMM should not be used without a precise inspection of the psychometric test's properties. In particular, a more adequate mixed model that accounts for curvilinearity and ceiling/floor effects should be fitted to evaluate the violation of the LMM assumptions and the reliability of the associations highlighted. For many psychometric tests, such as the MMSE and the BVRT, the standard LMM is most likely not reliable, and a mixed model that accounts for curvilinearity should be systematically preferred. The threshold mixed model is the most reliable model, since it models directly each level of the outcome, but it is computationally intensive. Therefore, for psychometric tests including a relatively large number of levels, a model assuming a curvilinear continuous outcome can be used instead. For example, the beta LMM (16) provided a fit similar to that of the threshold LMM and, more importantly, gave unbiased inference in our simulations while avoiding the computational problems. In contrast, for psychometric tests with a small number of levels, the threshold LMM should be preferred. The lcmm user-friendly R function is available within the lcmr R package (R Foundation for Statistical Computing, Vienna, Austria; http://cran.r-project.org/web/packages/lcmm/) for estimating latent process LMMs, including threshold and beta LMMs, and SAS macros are also available (SAS Institute Inc., Cary, North Carolina; http://biostat.isped.u-bordeaux2.fr).

The main concern regarding alternative mixed models is that, in contrast to the LMM, they do not provide an interpretation of risk-factor impact in terms of number of points lost on the test scale per time unit. This is actually a direct consequence of curvilinearity: The loss of 1 point on a test scale does not have the same meaning for the entire test range; it depends on the initial level. Thus, if degree of significance and the direction of association can be still interpreted as in the LMM, the intensity of association should rather be appreciated graphically as in Figure 4 or be quantified on the psychometric test scale in terms of the number of years a subject with a certain covariate value would need to reach the same cognitive level as a person with the covariate reference value (28). For example, in the illustration, a subject with a high educational level would reach the same MMSE score as a subject with a low educational level 14.7 years later at age 65 years and 12.9 years later at ages 70 and 75 years, respectively. In contrast, when using the standard LMM, the estimated effect corresponds to 8.1, 10.9, and 13.1 years at ages 65, 70, and 75 years, respectively.

In conclusion, to distinguish the impact of a covariate on the initial level of a scale from its impact on the change in quantitative scale scores over time (not only psychometric tests but also scales evaluating quality of life or activities of daily living), mixed models that account for their metrologic properties should be preferred over the LMM.

ACKNOWLEDGMENTS

Author affiliations: Biostatistics Department, INSERM U897, Institut de Santé Publique, d’Épidémiologie et de Développement, Université Bordeaux Segalen, Bordeaux, France (Cécile Proust-Lima, Hélène Jacqmin-Gadda); and Aging Department, INSERM U897, Institut de Santé Publique, d’Épidémiologie et de Développement, Université Bordeaux Segalen, Bordeaux, France (Jean-François Dartigues).

This work was carried out within the MOBIDYQ project (grant 2010 PRSP 006 01), which is funded by the Agence Nationale de la Recherche. The PAQUID Study was funded by Novartis Pharma AG, SCOR Insurance Ltd., Agrica, Conseil régional d’Aquitaine, Conseil Général de la Gironde, and Conseil Général de la Dordogne.

The authors thank Dr. Paul K. Crane for motivating discussions regarding the longitudinal analysis of psychometric tests.

Conflict of interest: none declared.

REFERENCES

APPENDIX

Computation of Akaike’s Information Criterion for discrete data

Using the latent process model formulation, the probability of observing the value \( k \) of an ordinal outcome \( Y \) with values in \( \{0, \ldots, C\} \) is

\[
P(Y_i = k) = P(\kappa_k \leq \Lambda(t_{ij}) + \epsilon_{ij} < \kappa_{k+1}) \quad \text{with} \quad \kappa = -\infty \quad \text{and} \quad \kappa_{C+1} = +\infty, \tag{A1}
\]

for subject \( i (i = 1, \ldots, n) \) and occasion \( j (j = 1, \ldots, n_i) \).

In a threshold model that considers an ordinal outcome, the thresholds \( \kappa_k \) for \( k = 1, \ldots, C \) are directly estimated. To compute the likelihood for the ordinal outcome using the other estimated mixed models that consider \( Y \) as continuous, equation A1 was also used with thresholds \( \kappa_k \) (\( k = 1, \ldots, C \)), defined as \( \kappa_k = k + 0.5 \) in the standard LMM, \( \kappa_k = \left[h(k) + h(k + 1)\right]/2 \) in the pretransformed LMM, and \( \kappa_k = \left[h(k, \hat{\eta}) + h(k + 1, \hat{\eta})\right]/2 \) in the beta LMM (with \( \hat{\eta} \) representing the maximum likelihood estimate of \( \eta \)).

Design of the simulation study

Samples were simulated to mimic the PAQUID cohort. Entry into the cohort was simulated according to a uniform distribution between ages 65 and 80 years for the general aging sample and 7–14 years before diagnosis for the prediagnostic phase of dementia sample, while dropout was simulated according to a uniform distribution between ages 80 and 95 years for the general aging sample and fixed at time 0 for the prediagnostic phase of dementia sample. In this window, each subject had a measurement every 3 years. The binary covariate (\( X_1 \) or \( X_2 \)) was simulated according to a Bernoulli distribution with a probability of 0.5.