British Journal of Mathematical and Statistical Psychology (2013), 66, 353–381

The British Psychological Society

www.wileyonlinelibrary.com

Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model

Sun-Joo Cho*, Michele Athay and Kristopher J. Preacher

Peabody College of Vanderbilt University, USA

Even though many educational and psychological tests are known to be multidimensional, little research has been done to address how to measure individual differences in change within an item response theory framework. In this paper, we suggest a generalized explanatory longitudinal item response model to measure individual differences in change. New longitudinal models for multidimensional tests and existing models for unidimensional tests are presented within this framework and implemented with software developed for generalized linear models. In addition to the measurement of change, the longitudinal models we present can also be used to explain individual differences in change scores for person groups (e.g., learning disabled students versus non-learning disabled students) and to model differences in item difficulties across item groups (e.g., number operation, measurement, and representation item groups in a mathematics test). An empirical example illustrates the use of the various models for measuring individual differences in change when there are person groups and multiple skill domains which lead to multidimensionality at a time point.

1. Introduction

In an item response theory (IRT) framework, longitudinal item response models (Andersen, 1985; Embretson, 1991; Fischer, 1976; te Marvelde, Glas, Landeghem, & Van Damme, 2006) can be used to measure individual differences in change over time. A common assumption of these item response models is that tests are unidimensional. That is, all dependencies in item responses are assumed to be explained using the time dimension[1] of the models. If a test is multidimensional at a time point, using models designed for unidimensional tests may not produce valid change estimates. This is mainly because all dependencies in item responses cannot be completely explained by the time dimension in longitudinal item response models (e.g., Andersen, 1985; Embretson, 1991).

*Correspondence should be addressed to Dr. Sun-Joo Cho, Peabody College of Vanderbilt University, 230 Appleton Place, Nashville, TN 37203-5721, USA (e-mail: sj.cho@vanderbilt.edu).

[1] We use the term 'time dimension' (denoted by θ_{jt} for a person j and a time point t) to describe the individual differences at each time point in a longitudinal data set and to indicate time as a source of individual differences.

DOI:10.1111/j.2044-8317.2012.02058.x

Along with the measurement of individual differences in change, the person group differences are often of interest. In educational measurement applications, for example, many research questions centre around person group difference information, such as differences across gender, prior instruction, learning style, and learning attribute groups (e.g., Cho, Bottge, Cohen, & Kim, 2011). In addition to person group information in longitudinal data, item group information can help the researcher to interpret the item parameters in an objective way. Explaining the item parameters (e.g., item difficulties) using the item group information provides internal validity to the test (Embretson, 1983). In addition, number operation, measurement, and representation item groups in a mathematics test can be linked to change scores in longitudinal item response models (Embretson, 1995).

In this study we introduce and develop *generalized explanatory longitudinal item response models* for multidimensional tests. These models provide information on person group differences in initial scores and change, individual differences in initial scores and change, item group differences in item difficulties, and item differences in item difficulties. The models we describe are illustrated with an empirical application using the lmer function (Bates & Maechler, 2009) from the lme4 package in R (R Development Core Team, 2012).

2. Motivating example: A multidimensional mathematics test

A test was created to measure the effects of the instructional intervention called Fraction of the Cost Instruction (FOC) on three specific skill domains: number operation, measurement, and representation skills. The same FOC items were given to low-achieving adolescents, including learning disabled (LD) students, and to non-LD students over three time points. A researcher who developed the FOC test was interested in answering the following research questions:

1. Is there any change score difference between LD students and non-LD students?
2. Can we interpret the change scores based on relative complexity characterized by item difficulties across the different cognitive skills?
3. How much progress do students make after intervention?

2.1. Fraction of the cost

The FOC intervention began with an 8-minute video (Bottge, Heinrichs, Chan, Mehta, & Watson, 2003) starring three middle school students in a small Midwestern town in the US. Available in both Spanish and English, the video served to introduce viewers to the problem of building a skateboard ramp. Several interrelated skills/tasks were then required. First, students were required to learn the mathematics needed to solve several real-world problem-based tasks related to building the ramp. Next, students had to learn how to read schematic plans or blueprints to determine the size and quantity of materials (wood, screws, etc.) required to construct the ramp. Finally, after determining how best to cut the wood with the least waste possible, students were then asked to build the ramp.

2.2. FOC mathematics test

To assess the effects of the FOC instruction, a test was constructed based on standards set by the US National Council of Teachers of Mathematics for students in Grades 6–8.

Consisting of 23 items, the test used in this study contained performance task items assessing the abilities of students to interpret a three-dimensional schematic plan, measure lengths of building materials in feet and inches, estimate and compute combinations using whole numbers and fractions, and interpret and record data in tables. Of the 23 items, 20 items asked students to respond with short answers. These were scored as either correct or incorrect. However, six of these items displayed low or zero proportions of correct responses at the first two time points. Therefore, these items were dropped from the analysis. The result was a 14-item measure. Table 1 presents a matrix showing the particular mathematical skills required to solve each item.

2.3. Samples and multi-wave study design

The sample consisted of 109 Grade 7 students (46% male) drawn from six maths classes in a small Midwestern school district in the US. Nine students were diagnosed as LD students.

The study was conducted from October to April in the same school year. During these seven school months, the FOC test was administered three separate times: time point 1 (pre-test 1) occurred in October (109 students); time point 2 (pre-test 2) occurred in November (107 students); and time point 3 (post-test) occurred after the FOC instruction in April (109 students).

2.4. Multidimensionality of the test

Multidimensionality in item response models can be considered to capture heterogeneity that might arise due to item content, when the various dimensions of a test are measured by different subgroups of items. De Boeck (2011) found that the change depends on these three skill domains using the FOC test, indicating that the FOC test is multidimensional due to item groups. Each skill domain is assumed to be unidimensional, following De Boeck (2011).

Table 1. Cognitive skills required for solving each item

	Cognitive skills		
Item	Number operation	Measurement	Representation
1	1	0	0
2	1	0	0
3	1	1	0
4	1	1	0
5	1	1	0
6	1	1	0
7	1	1	0
8	1	1	0
9	0	1	0
10	1	1	1
11	1	1	1
12	1	0	0
13	1	0	0
14	1	0	0

When person group differences on a dimension are investigated such as in research question 1, it is assumed that the test measures the dimension in the same way across all person groups being compared. This is often called measurement invariance. Measurement invariance is important because it is not possible to measure reliable and valid change such as in research questions 2 and 3 unless there is evidence that the same dimension is being measured over time. For longitudinal data having multiple person groups, there are at least four ways in which the assumption of measurement invariance can be violated: item parameters differ (1) across person groups; (2) across time points; (3) across person groups within a time point; or (4) across time points within a person group. The investigation of all four of these violations of measurement invariance can be done using differential item functioning (DIF) analysis. DIF for the person groups means that the same levels of the dimension (e.g., ability) lead to different response probabilities depending on the person group to which one belongs (Scheuneman, 1979). DIF for time points indicates that the same levels of the dimension (e.g., ability) lead to different response probabilities depending on a time point.

In our example, DIF across LD and non-LD groups at each time point was investigated initially to make sure that the dimension being measured is the same across the groups at each time point. To measure change, DIF across time points was investigated to check if the dimension (as measured by the same set of items) was the same over the three time points. Total-score based DIF procedures such as Mantel–Haenszel statistics, the standardization method, and logistic regression models were implemented across LD and non-LD groups at each time point. Two items (items 10 and 11 in Table 1) were detected as DIF items, but the effect sizes of the DIF magnitude were small. We concluded that the same dimension was measured across LD and non-LD groups at each time point. Likelihood ratio tests using a Rasch model and the three total-score based DIF procedures used for a person group were then implemented between three pairs of three time points. Three items (items 11–13 in Table 1) were detected as DIF items based on the likelihood ratio procedure, but the effect size of DIF magnitudes was moderate. No DIF items were found based on the three total-score based DIF items. Based on these DIF analyses, it can be concluded that there is evidence that the same dimension was measured across person groups and time points.

3. Explanatory longitudinal item response models

When a test is multidimensional due to the existence of groups of items (e.g., different skill sets), change is measured within each item group because item responses would be conditionally independent within an item group d over time points. With this conditional independence assumption, the time parameter or dimension (denoted by θ_{jtd} for a person j, a time point t, and an item group d) is included in the model for each item group, not for all items (one group of items). For the three time points and an item group 1 ($d = 1$), as an example, the time parameters are θ_{j11}, θ_{j21}, and θ_{j31} for each time point, respectively. The multidimensionality of the three time parameters ($\boldsymbol{\theta}_{j1} = [\theta_{j11}, \theta_{j21}, \theta_{j31}]'$) is considered to measure individual differences in change within the item group.

Along with having the individual differences in change and individual item difficulties in the model as a purely descriptive approach, they can be modelled as functions of various covariates in an explanatory approach (De Boeck & Wilson, 2004). The effects of these covariates can be considered as averages, where the differences between individuals or item difficulties have been averaged over. Individual differences in change can be explained by three covariate effects: (1) the time covariate effect as average score

for each time point; (2) the person group covariate effect as average score for each group; and (3) the interaction (between time point and person group) covariate effect as average score for each time point and each person group. Individual item difficulties can be explained by item group as the item covariate. The item group difference information can be linked to the change score interpretation (Embretson, 1995). When item groups are used to explain item parameters in item response models, it is preferable to allow *random* residuals across items because item group information may not explain all variations in item parameters. Ignoring the random residuals leads to underestimated standard errors of the estimate for the item group difference (Janssen, Schepers, & Peres, 2004).

An explanatory longitudinal item response model is given by

$$\text{logit}[P(y_{g[j]d[i]t} = 1 \mid \theta_{jtd}, \epsilon_{d[i]})] = \zeta_{..} + (\zeta_{tg} + \zeta_{t.} + \zeta_{.g} + \theta_{jtd}) - (\beta_d + \epsilon_{d[i]}), \qquad (1)$$

where

- i is an index for an item ($i = 1,\ldots,I$),
- j is an index for a person ($j = 1,\ldots,J$),
- t is an index for a time point ($t = 1,\ldots,T$),
- g is an index for a person group ($g = 1,\ldots,G$),
- d is an index for an item group ($d = 1,\ldots,D$),
- $d[i]$ indicates that an item i is nested within an item group d,
- $g[j]$ indicates that a person j is nested within a person group g,
- $\zeta_{..}$ is a fixed intercept parameter (average score across persons, items, and time points),
- ζ_{tg} is a fixed interaction parameter of time and the person group effect (average score for each time point and each person group),
- $\zeta_{t.}$ is a fixed time parameter (average score for each time point),
- $\zeta_{.g}$ is a fixed person group parameter (average score for each person group),
- β_d is a fixed item difficulty parameter for an item group (or feature) d (average difficulty for each item group),
- θ_{jtd} is a time (change) parameter or dimension, random across persons, and
- $\epsilon_{d[i]}$ is an item difficulty parameter, random across items.

$\boldsymbol{\theta}_{jd}$ is assumed to follow a multivariate normal (MN) distribution for each item group d as follows: $\boldsymbol{\theta}_{jd} = [\theta_{j1d}, \ldots, \theta_{jtd}, \ldots, \theta_{jTd}]' \sim MN(\boldsymbol{\mu}_{d(T\times 1)}, \Sigma_{d(T\times T)})$. $\epsilon_{d[i]}$ is assumed to follow a normal (N) distribution: $\epsilon_{d[i]} \sim N(\mu_\epsilon, \sigma_\epsilon)$. The means of random effects, $\boldsymbol{\mu}_{d(T\times 1)}$ and μ_ϵ, are set to $\mathbf{0}$ to identify the model.

3.1. Illustration of the model using the FOC test data

The model equation (1) can be used to answer the previously stated research questions. Parameters were estimated using the R lmer function and the results are shown in Table 2. Indicators were used for time points (Qs as shown in Table 4), person groups (0 for non-LD students and 1 for LD students), and item groups (as shown in Table 1).

For the first research question, the two parameter estimates ($\hat{\zeta}_{tg}$ and $\hat{\zeta}_{.g}$) were interpreted. The two interaction effects, time 2 $\times$ group and time 3 $\times$ group ($\hat{\zeta}_{21} = -0.147$ and $\hat{\zeta}_{31} = 0.681$, respectively), are not significant, indicating that the effect of time is the same for both the LD and the non-LD student groups when other

Table 2. Estimates (Est.) and standard errors (SEs) from the explanatory longitudinal item response model

		Est.			SEs
Fixed effects					
$\zeta_{..}$ [Intercept]		0.608			0.946
Persons					
ζ_{21} [Time 2 by Group]		−0.147			0.559
ζ_{31} [Time 3 by Group]		0.681			0.594
$\zeta_{2.}$ [Time 2]		0.694*			0.151
$\zeta_{3.}$ [Time 3]		2.546*			0.173
$\zeta_{.1}$ [Group]		−1.121*			0.485
Items					
β_1 [Operation]		−1.638*			0.876
β_2 [Measurement]		0.369			0.501
β_3 [Representation]		−0.631			0.672
Random effects					
Persons					
		Var.	Corr.		
			Time 1	Time 2	
Σ_1 [Operation]	Time 1	3.132			
	Time 2	0.454	1.000		
	Time 3	2.405	−1.000	1.000	
Σ_2 [Measurement]	Time 1	2.557			
	Time 2	3.015	−0.148		
	Time 3	3.243	−0.423	−0.748	
Σ_3 [Representation]	Time 1	1.137			
	Time 2	0.865	−0.464		
	Time 3	0.505	−0.439	0.154	
Items					
σ_ϵ [Variance]		0.624			

Note. *Significance based on p-value <.05 for fixed effects.

covariates are 0. A change in ability over time ($\hat{\zeta}_{2.} = 0.694$ and $\hat{\zeta}_{3.} = 2.546$) is significant at time 2 and time 3 when other covariates are 0. A significant time effect at time 3 shows the intervention effect. There is a significant person (i.e., student) group difference for $\hat{\zeta}_{.1}$ whereby the overall ability across time points for the LD students is 1.121 lower on the logit scale than that of the non-LD students when other covariates are 0.

For the second research question, the person scores can be connected with item complexity represented by $\hat{\beta}_d$. In the model, the difficulties of the item skill domains are modelled instead of the difficulty of each item. The difficulty order indicates that the measurement domain is the hardest one, followed by the representation skill and the operation skill. The coefficient for the operation skill was statistically significant. One can map the initial and change scores to the item complexity with respect to item group difficulty, $\hat{\beta}_d$. For example, if a student's score at time 3 (initial score + change score at time 2 + change score at time 3) is higher than the *measurement* item difficulty ($\hat{\beta}_2$), the probability of getting a correct response for *measurement* items is higher than 0.5 at time 3 after the intervention. The moderate residual variance estimate ($\hat{\sigma}_\epsilon = 0.624$) shows that a large portion of the residuals remains, indicating that there is still variation in item difficulties after explaining them with item skill domains. When the residual variance is ignored, standard errors of $\hat{\beta}_d$ are seriously underestimated.

For the third research question, an initial and change score for each person at time 2 and time 3 were predicted based on $\hat{\Sigma}_{d(T\times T)}$. The scores can be considered as residual person scores for the individual differences in initial and change scores after they are explained by the interaction of time and group covariate effect ($\hat{\zeta}_{tg}$), time covariate effect ($\hat{\zeta}_{t\cdot}$), and person group covariate effect ($\hat{\zeta}_{\cdot g}$). To obtain the initial and change scores, the person and time fixed effects ($\hat{\zeta}_{tg} + \hat{\zeta}_{t\cdot} + \hat{\zeta}_{\cdot g}$) were added to the residual person scores. The scores at time 1 (before intervention) and time 3 (after intervention) for each skill domain are shown in Figure 1. Scores at time 1 are initial scores and scores at time 3 (initial score + change score at time 2 + change score at time 3) are scores after intervention. As shown in Figure 1, for number operation and measurement skills, students who were at the low score level at time 1 tend to have the higher scores at time 3 after intervention. However, students who were at the high score level at time 1 appear to have the lower scores at time 3 after intervention. One possible interpretation of this pattern is that more able students were losing motivation to solve the same items over three time points. The perfect correlation between scores at time 1 and time 3 for the number operation skill may be due to an estimation error, as we will discuss in the discussion section. For the representation skill, there are more students who shifted from the lower level at time 1 to the higher level at time 3 after intervention.

4. A generalized explanatory longitudinal item response model

A generalized explanatory longitudinal item response model is presented to show how the comparison models can be specified using one modelling framework so that the similarities and differences between models can be shown. In addition, the model specification with the generalized explanatory longitudinal item response model is directly linked to R lmer syntax.

Indicators are added to each parameter of an explanatory longitudinal item response model (equation (1)) in the generalized explanatory longitudinal item response model given by

$$\begin{aligned}\text{logit}[P(y_{g[j]d[i]t} = 1|\theta_{jtd}, \epsilon_{d[i]})] = \zeta_{\cdot\cdot}Z_{0j}X_{0i} + (\zeta_{tg}Q_{d[i]t}Z_{g[j]} + \zeta_{t\cdot}Q_{d[i]t} \\ + \zeta_{\cdot g}Z_{g[j]} + \theta_{jtd}Q_{d[i]t}) - \left(\beta_d X_{d[i]} + \epsilon_{d[i]}X_{0i}\right),\end{aligned} \tag{2}$$

where

- Z_{0j} is an indicator for a person j ($Z_{0j} = 1$ for all js),
- X_{0i} is an indicator for an item i ($X_{0i} = 1$ for all is),
- $Z_{g[j]}$ is an indicator for a person j nested within a person group g,
- $Q_{d[i]t}$ is an indicator for an item i nested within an item group d and a time point t, and
- $X_{d[i]}$ is an indicator for an item i nested within an item group d.

The model fit of the explanatory longitudinal item response model for the multidimensional tests can be compared with that of descriptive longitudinal item response models (without covariates) to investigate the effects of covariates and that of longitudinal models for unidimensional tests (Andersen, 1985; Embretson, 1991) as comparison models with respect to dimensionality. The Embretson (1991) model was chosen for the comparison because it provides the direct change from the second time point. Specifically, the following four models were compared:

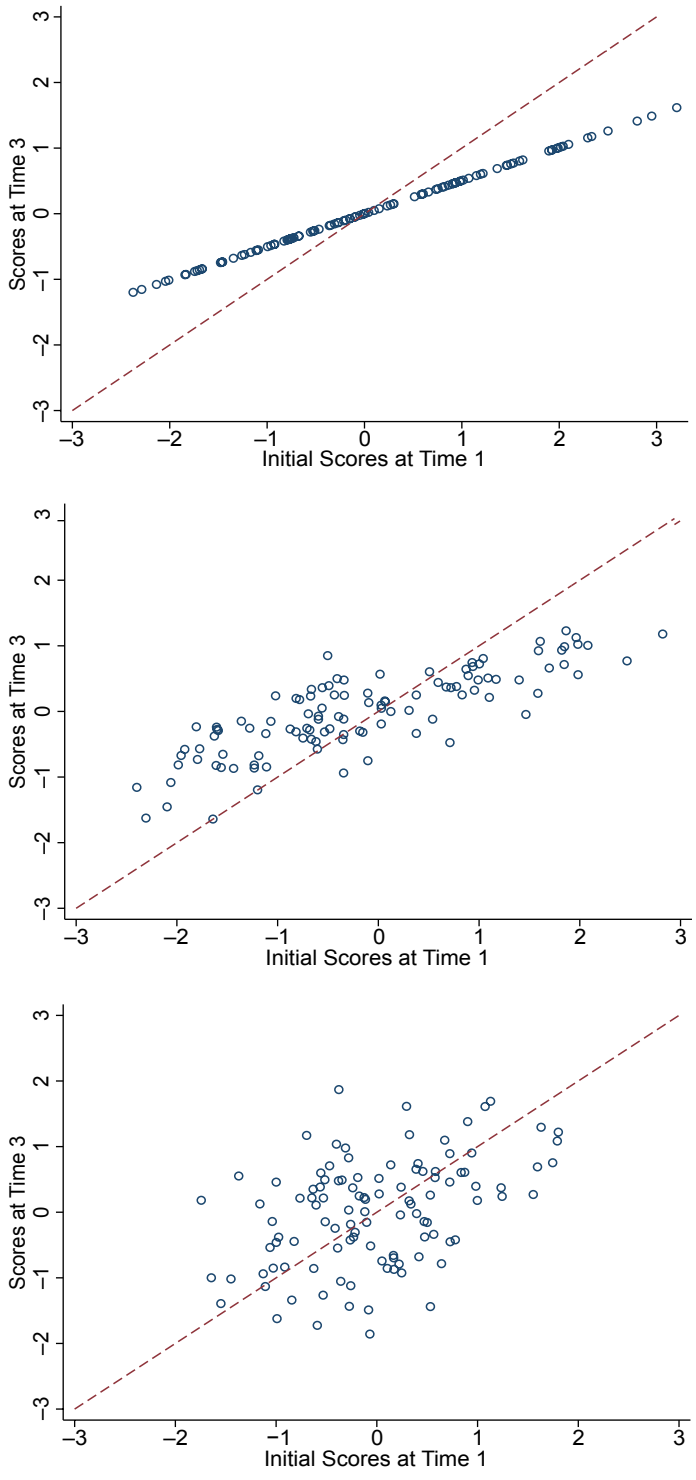

Figure 1. Scores for number operation skill (top); scores for measurement skill (middle); scores for representation skill (bottom).

- model 1, Embretson descriptive model for unidimensional tests;
- model 2, Embretson explanatory model for unidimensional tests;
- model 3, descriptive longitudinal model for multidimensional tests (without fixed effects in equation (2));
- model 4, explanatory longitudinal model for multidimensional tests (equation (2)).

The following section shows how these four models can be specified as special cases of the generalized explanatory longitudinal item response model to see the similarities and differences between the longitudinal models for unidimensional and multidimensional tests. The specification as a general model can be used as the data structure in the R lmer function as shown in Appendix B.

4.1. Illustration of models using the FOC test data

$Q_{d[i]t}$ is an indicator specifying the time dimension for change in the model. Different kinds of longitudinal item response models can be described using different $Q_{d[i]t}$ specifications in equation (2). Below, the specification of $Q_{d[i]t}$ in the data structure is described for models 1 and 2 first and for models 3 and 4 subsequently for the FOC data.

Table 3 shows the indicators for models 1 and 2 for the FOC data with three time points, 14 items, and two persons. Two persons were chosen for the sake of simplicity. Embretson (1991) used a Wiener simplex process to obtain the initial status and change scores for each time point. The three-time-point example of the Wiener simplex process (3×3 matrix) is

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix},$$

where rows indicate the time points and columns indicate the dimension. An element of the matrix is 1 if a dimension is modelled in that time point and 0 otherwise. The structure shows that performance depends on initial status for time 1, initial status and change at time 2 for time 2, and initial status, change at time 2, and change at time 3 for time 3.

The Wiener simplex structure can be specified with $Q_{d[i]t}$ in longitudinal item response models. Following Embretson (1991), the FOC data are assumed to be unidimensional, which means that all items are from a single item group ($d = 1$). As shown in Table 3, the Wiener simplex applies to the structure of $Q_{1[i]t}$ such that $Q_{1[i]1}$ (indicator for time 1) is 1 for all three time points t, $Q_{1[i]2}$ (indicator for time 2) is 0 for time 1 and 1 for time 2 and time 3, and $Q_{1[i]3}$ (indicator for time 3) is 0 for time 1 and time 2 and 1 for time 3. For model 1 as a descriptive model, the indicator for items ($X_{1[i]}$) is specified for each item, and the same indicator across three time points is specified to impose measurement invariance. Because the variance in item difficulty (σ_ϵ) is assumed to be explained by the difficulty of all individual items, it is 0 in the descriptive model. For model 2 as an explanatory model, the indicator for items ($X_{1[i]}, X_{2[i]}$, and $X_{3[i]}$) is specified with respect to three item groups ($d = 1,2,3$). $X_{1[i]}$ is 1 if items measure the number operation skill and 0 otherwise, $X_{2[i]}$ is 1 if items measure the measurement skill and 0 otherwise, and $X_{3[i]}$ is 1 if items measure the representation skill and 0 otherwise. Because item groups may not explain variation in item difficulty across items, the random effect ϵ_i is allowed to be estimated for the residual. In model 2, the indicator for persons ($Z_{g[j]}$ where g is 1) is 1 if students were classified in the LD group and 0 otherwise.

Table 3. Indicators of the longitudinal item response model for unidimensional tests

							Model 1					Model 2					
t	j	i	$y_{g[j]d[i]t}$	$Q_{1[i]1}$	$Q_{1[i]2}$	$Q_{1[i]3}$	$X_{1[1]}$	$X_{1[2]}$	…	$X_{1[13]}$	$X_{1[14]}$	$X_{1[i]}$	$X_{2[i]}$	$X_{3[i]}$	Z_{0j}	X_{0i}	$Z_{1[j]}$
1	1	1	1	1	0	0	1	0	…	0	0	1	0	0	1	1	0
1	2	1	0	1	0	0	1	0	…	0	0	1	0	0	1	1	1
1	1	2	0	1	0	0	0	1	…	0	0	1	0	0	1	1	0
1	2	2	0	1	0	0	0	1	…	0	0	1	0	0	1	1	1
1	1	3	0	1	0	0	0	0	…	0	0	1	1	0	1	1	0
1	2	3	0	1	0	0	0	0	…	0	0	1	1	0	1	1	1
1	1	4	0	1	0	0	0	0	…	0	0	1	1	0	1	1	0
1	2	4	0	1	0	0	0	0	…	0	0	1	1	0	1	1	1
1	1	5	1	1	0	0	0	0	…	0	0	1	1	0	1	1	0
1	2	5	0	1	0	0	0	0	…	0	0	1	1	0	1	1	1
1	1	6	0	1	0	0	0	0	…	0	0	1	1	0	1	1	0
1	2	6	0	1	0	0	0	0	…	0	0	1	1	0	1	1	1
1	1	7	0	1	0	0	0	0	…	0	0	1	1	0	1	1	0
1	2	7	0	1	0	0	0	0	…	0	0	1	1	0	1	1	1
1	1	8	0	1	0	0	0	0	…	0	0	1	1	0	1	1	0
1	2	8	0	1	0	0	0	0	…	0	0	1	1	0	1	1	1
1	1	9	0	1	0	0	0	0	…	0	0	0	1	0	1	1	0
1	2	9	0	1	0	0	0	0	…	0	0	0	1	0	1	1	1
1	1	10	0	1	0	0	0	0	…	0	0	1	1	1	1	1	0
1	2	10	0	1	0	0	0	0	…	0	0	1	1	1	1	1	1
1	1	11	0	1	0	0	0	0	…	0	0	1	1	1	1	1	0
1	2	11	0	1	0	0	0	0	…	0	0	1	0	1	1	1	1
1	1	12	1	1	0	0	0	0	…	0	0	1	0	0	1	1	0
1	2	12	0	1	0	0	0	0	…	0	0	1	0	0	1	1	1
1	1	13	1	1	0	0	0	0	…	1	0	1	0	0	1	1	0
1	2	13	0	1	0	0	0	0	…	1	0	1	0	0	1	1	1
1	1	14	0	1	0	0	0	0	…	0	1	1	0	0	1	1	0
1	2	14	0	1	0	0	0	0	…	0	1	1	0	0	1	1	1
2	1	1	0	1	1	0	1	0	…	0	0	1	0	0	1	1	0
2	2	1	0	1	1	0	1	0	…	0	0	1	0	0	1	1	1
2	1	2	0	1	1	0	0	1	…	0	0	1	0	0	1	1	0
2	2	2	0	1	1	0	0	1	…	0	0	1	0	0	1	1	1
2	1	3	0	1	1	0	0	0	…	0	0	1	1	0	1	1	0
2	2	3	0	1	1	0	0	0	…	0	0	1	1	0	1	1	1
2	1	4	0	1	1	0	0	0	…	0	0	1	1	0	1	1	0
2	2	4	0	1	1	0	0	0	…	0	0	1	1	0	1	1	1
2	1	5	1	1	1	0	0	0	…	0	0	1	1	0	1	1	0
2	2	5	0	1	1	0	0	0	…	0	0	1	1	0	1	1	1
2	1	6	0	1	1	0	0	0	…	0	0	1	1	0	1	1	0
2	2	6	0	1	1	0	0	0	…	0	0	1	1	0	1	1	1
2	1	7	0	1	1	0	0	0	…	0	0	1	1	0	1	1	0
2	2	7	0	1	1	0	0	0	…	0	0	1	1	0	1	1	1
2	1	8	0	1	1	0	0	0	…	0	0	1	1	0	1	1	0
2	2	8	0	1	1	0	0	0	…	0	0	1	1	0	1	1	1
2	1	9	0	1	1	0	0	0	…	0	0	0	1	0	1	1	0
2	2	9	0	1	1	0	0	0	…	0	0	0	1	0	1	1	1

Continued

Table 3. (*Continued*)

							Model 1					Model 2					
t	j	i	$y_{g[j]d[i]t}$	$Q_{1[i]1}$	$Q_{1[i]2}$	$Q_{1[i]3}$	$X_{1[1]}$	$X_{1[2]}$	…	$X_{1[13]}$	$X_{1[14]}$	$X_{1[i]}$	$X_{2[i]}$	$X_{3[i]}$	Z_{0j}	X_{0i}	$Z_{1[j]}$
2	1	10	0	1	1	0	0	0	…	0	0	1	1	1	1	1	0
2	2	10	0	1	1	0	0	0	…	0	0	1	1	1	1	1	1
2	1	11	0	1	1	0	0	0	…	0	0	1	1	1	1	1	0
2	2	11	0	1	1	0	0	0	…	0	0	1	0	1	1	1	1
2	1	12	1	1	1	0	0	0	…	0	0	1	0	0	1	1	0
2	2	12	1	1	1	0	0	0	…	0	0	1	0	0	1	1	1
2	1	13	1	1	1	0	0	0	…	1	0	1	0	0	1	1	0
2	2	13	0	1	1	0	0	0	…	1	0	1	0	0	1	1	1
2	1	14	0	1	1	0	0	0	…	0	1	1	0	0	1	1	0
2	2	14	0	1	1	0	0	0	…	0	1	1	0	0	1	1	1
3	1	1	0	1	1	1	1	0	…	0	0	1	0	0	1	1	0
3	2	1	0	1	1	1	1	0	…	0	0	1	0	0	1	1	1
3	1	2	0	1	1	1	0	1	…	0	0	1	0	0	1	1	0
3	2	2	0	1	1	1	0	1	…	0	0	1	0	0	1	1	1
3	1	3	0	1	1	1	0	0	…	0	0	1	1	0	1	1	0
3	2	3	0	1	1	1	0	0	…	0	0	1	1	0	1	1	1
3	1	4	0	1	1	1	0	0	…	0	0	1	1	0	1	1	0
3	2	4	0	1	1	1	0	0	…	0	0	1	1	0	1	1	1
3	1	5	0	1	1	1	0	0	…	0	0	1	1	0	1	1	0
3	2	5	1	1	1	1	0	0	…	0	0	1	1	0	1	1	1
3	1	6	0	1	1	1	0	0	…	0	0	1	1	0	1	1	0
3	2	6	0	1	1	1	0	0	…	0	0	1	1	0	1	1	1
3	1	7	0	1	1	1	0	0	…	0	0	1	1	0	1	1	0
3	2	7	0	1	1	1	0	0	…	0	0	1	1	0	1	1	1
3	1	8	0	1	1	1	0	0	…	0	0	1	1	0	1	1	0
3	2	8	0	1	1	1	0	0	…	0	0	1	1	0	1	1	1
3	1	9	0	1	1	1	0	0	…	0	0	0	1	0	1	1	0
3	2	9	0	1	1	1	0	0	…	0	0	0	1	0	1	1	1
3	1	10	0	1	1	1	0	0	…	0	0	1	1	1	1	1	0
3	2	10	0	1	1	1	0	0	…	0	0	1	1	1	1	1	1
3	1	11	0	1	1	1	0	0	…	0	0	1	1	1	1	1	0
3	2	11	0	1	1	1	0	0	…	0	0	1	0	1	1	1	1
3	1	12	0	1	1	1	0	0	…	0	0	1	0	0	1	1	0
3	2	12	0	1	1	1	0	0	…	0	0	1	0	0	1	1	1
3	1	13	0	1	1	1	0	0	…	1	0	1	0	0	1	1	0
3	2	13	0	1	1	1	0	0	…	1	0	1	0	0	1	1	1
3	1	14	0	1	1	1	0	0	…	0	1	1	0	0	1	1	0
3	2	14	0	1	1	1	0	0	…	0	1	1	0	0	1	1	1

Table 4 shows the indicators for models 3 and 4 for three time points, 14 items, and two persons. For the multidimensional tests, the Wiener simplex applies to the structure of $Q_{d[i]t}$ for each item group d (d = 1,2,3). $Q_{d[i]t}$ can be specified $Q_{1[i]1}$, $Q_{1[i]2}$, and $Q_{1[i]3}$ for number operation items; $Q_{2[i]1}$, $Q_{2[i]2}$, and $Q_{2[i]3}$ for measurement items; and $Q_{3[i]1}$, $Q_{3[i]2}$, and $Q_{3[i]3}$ for representation items for times 1, 2, and 3, respectively. One can create these nine indicators for the Wiener simplex easily by multiplying $Q_{d[i]t}$ by $X_{d[i]}$ specified in Table 3. For example, $Q_{1[i]1}$, $Q_{1[i]2}$, and $Q_{1[i]3}$ in Table 4 are from $Q_{1[i]1} \times X_{1[i]}$, $Q_{1[i]2} \times X_{2[i]}$, and $Q_{1[i]3} \times X_{3[i]}$ in Table 3, respectively. For model 3 as a descriptive

Table 4. Indicators of the longitudinal item response model for multidimensional tests

t	j	i	$y_{g[j]d[i]t}$	$Q_{0[i]1}$	$Q_{0[i]2}$	$Q_{0[i]3}$	$Q_{1[i]1}$	$Q_{1[i]2}$	$Q_{1[i]3}$	$Q_{2[i]1}$	$Q_{2[i]2}$	$Q_{2[i]3}$	$Q_{3[i]1}$	$Q_{3[i]2}$	$Q_{3[i]3}$
1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0
1	2	1	0	1	0	0	1	0	0	0	0	0	0	0	0
1	1	2	0	1	0	0	1	0	0	0	0	0	0	0	0
1	2	2	0	1	0	0	1	0	0	0	0	0	0	0	0
1	1	3	0	1	0	0	1	0	0	1	0	0	0	0	0
1	2	3	0	1	0	0	1	0	0	1	0	0	0	0	0
1	1	4	0	1	0	0	1	0	0	1	0	0	0	0	0
1	2	4	0	1	0	0	1	0	0	1	0	0	0	0	0
1	1	5	1	1	0	0	1	0	0	1	0	0	0	0	0
1	2	5	0	1	0	0	1	0	0	1	0	0	0	0	0
1	1	6	0	1	0	0	1	0	0	1	0	0	0	0	0
1	2	6	0	1	0	0	1	0	0	1	0	0	0	0	0
1	1	7	0	1	0	0	1	0	0	1	0	0	0	0	0
1	2	7	0	1	0	0	1	0	0	1	0	0	0	0	0
1	1	8	0	1	0	0	1	0	0	1	0	0	0	0	0
1	2	8	0	1	0	0	1	0	0	1	0	0	0	0	0
1	1	9	0	1	0	0	0	0	0	1	0	0	0	0	0
1	2	9	0	1	0	0	0	0	0	1	0	0	0	0	0
1	1	10	0	1	0	0	1	0	0	1	0	0	1	0	0
1	2	10	0	1	0	0	1	0	0	1	0	0	1	0	0
1	1	11	0	1	0	0	1	0	0	1	0	0	1	0	0
1	2	11	0	1	0	0	1	0	0	1	0	0	1	0	0
1	1	12	1	1	0	0	1	0	0	0	0	0	0	0	0
1	2	12	0	1	0	0	1	0	0	0	0	0	0	0	0
1	1	13	1	1	0	0	1	0	0	0	0	0	0	0	0
1	2	13	0	1	0	0	1	0	0	0	0	0	0	0	0
1	1	14	0	1	0	0	1	0	0	0	0	0	0	0	0
1	2	14	0	1	0	0	1	0	0	0	0	0	0	0	0
2	1	1	0	1	1	0	1	1	0	0	0	0	0	0	0
2	2	1	0	1	1	0	1	1	0	0	0	0	0	0	0
2	1	2	0	1	1	0	1	1	0	0	0	0	0	0	0
2	2	2	0	1	1	0	1	1	0	0	0	0	0	0	0
2	1	3	0	1	1	0	1	1	0	1	1	0	0	0	0
2	2	3	0	1	1	0	1	1	0	1	1	0	0	0	0
2	1	4	0	1	1	0	1	1	0	1	1	0	0	0	0
2	2	4	0	1	1	0	1	1	0	1	1	0	0	0	0
2	1	5	1	1	1	0	1	1	0	1	1	0	0	0	0
2	2	5	0	1	1	0	1	1	0	1	1	0	0	0	0
2	1	6	0	1	1	0	1	1	0	1	1	0	0	0	0
2	2	6	0	1	1	0	1	1	0	1	1	0	0	0	0
2	1	7	0	1	1	0	1	1	0	1	1	0	0	0	0
2	2	7	0	1	1	0	1	1	0	1	1	0	0	0	0
2	1	8	0	1	1	0	1	1	0	1	1	0	0	0	0
2	2	8	0	1	1	0	1	1	0	1	1	0	0	0	0
2	1	9	0	1	1	0	0	0	0	1	1	0	0	0	0
2	2	9	0	1	1	0	0	0	0	1	1	0	0	0	0
2	1	10	0	1	1	0	1	1	0	1	1	0	1	1	0
2	2	10	0	1	1	0	1	1	0	1	1	0	1	1	0

Continued

Table 4. (*Continued*)

t	j	i	$y_{g[j]d[i]t}$	$Q_{0[i]1}$	$Q_{0[i]2}$	$Q_{0[i]3}$	$Q_{1[i]1}$	$Q_{1[i]2}$	$Q_{1[i]3}$	$Q_{2[i]1}$	$Q_{2[i]2}$	$Q_{2[i]3}$	$Q_{3[i]1}$	$Q_{3[i]2}$	$Q_{3[i]3}$
2	1	11	0	1	1	0	1	1	0	1	1	0	1	1	0
2	2	11	0	1	1	0	1	1	0	1	1	0	1	1	0
2	1	12	1	1	1	0	1	1	0	0	0	0	0	0	0
2	2	12	1	1	1	0	1	1	0	0	0	0	0	0	0
2	1	13	1	1	1	0	1	1	0	0	0	0	0	0	0
2	2	13	0	1	1	0	1	1	0	0	0	0	0	0	0
2	1	14	0	1	1	0	1	1	0	0	0	0	0	0	0
2	2	14	0	1	1	0	1	1	0	0	0	0	0	0	0
3	1	1	0	1	1	1	1	1	1	0	0	0	0	0	0
3	2	1	0	1	1	1	1	1	1	0	0	0	0	0	0
3	1	2	0	1	1	1	1	1	1	0	0	0	0	0	0
3	2	2	0	1	1	1	1	1	1	0	0	0	0	0	0
3	1	3	0	1	1	1	1	1	1	1	1	1	0	0	0
3	2	3	0	1	1	1	1	1	1	1	1	1	0	0	0
3	1	4	0	1	1	1	1	1	1	1	1	1	0	0	0
3	2	4	0	1	1	1	1	1	1	1	1	1	0	0	0
3	1	5	0	1	1	1	1	1	1	1	1	1	0	0	0
3	2	5	1	1	1	1	1	1	1	1	1	1	0	0	0
3	1	6	0	1	1	1	1	1	1	1	1	1	0	0	0
3	2	6	0	1	1	1	1	1	1	1	1	1	0	0	0
3	1	7	0	1	1	1	1	1	1	1	1	1	0	0	0
3	2	7	0	1	1	1	1	1	1	1	1	1	0	0	0
3	1	8	0	1	1	1	1	1	1	1	1	1	0	0	0
3	2	8	0	1	1	1	1	1	1	1	1	1	0	0	0
3	1	9	0	1	1	1	0	0	0	1	1	1	0	0	0
3	2	9	0	1	1	1	0	0	0	1	1	1	0	0	0
3	1	10	0	1	1	1	1	1	1	1	1	1	1	1	1
3	2	10	0	1	1	1	1	1	1	1	1	1	1	1	1
3	1	11	0	1	1	1	1	1	1	1	1	1	1	1	1
3	2	11	0	1	1	1	1	1	1	1	1	1	1	1	1
3	1	12	0	1	1	1	1	1	1	0	0	0	0	0	0
3	2	12	0	1	1	1	1	1	1	0	0	0	0	0	0
3	1	13	0	1	1	1	1	1	1	0	0	0	0	0	0
3	2	13	0	1	1	1	1	1	1	0	0	0	0	0	0
3	1	14	0	1	1	1	1	1	1	0	0	0	0	0	0
3	2	14	0	1	1	1	1	1	1	0	0	0	0	0	0

model, the same indicator for items ($X_{1[i]}$) in model 1 can be set. For model 4 as an explanatory model, the same indicators for items and persons ($X_{1[i]}, X_{2[i]}, X_{3[i]}$, and $Z_{1[j]}$) in model 2 can be specified.

In explanatory models (models 2 and 4), $Z_{0j}X_{0i}$ is an indicator for the intercept, $\zeta_{..}$. (See Table 3 for the specification of Z_{0j} and X_{0i}.) The lmer code for all four models is shown in Appendix A. The data structure shown in Tables 3 and 4 is matched with the R lmer code in Appendix B.

Table 5 shows the Akaike and Bayesian information criteria, AIC (Akaike, 1974) and BIC (Schwarz, 1978), for models 1–4. Multidimensional models (models 3 and 4) fit better than unidimensional models (models 1 and 2), showing evidence for

Table 5. Model fit comparisons for the model to measure change for each skill domain

Model	Dimension	Explanatory	No. of parameters	Log-likelihood	AIC	BIC
Model 1	Unidimensional	No	20	−2462.9	4965.8	5094.4
Model 2	Unidimensional	Yes	16	−2413.7	4859.4	4962.3
Model 3	Multidimensional	No	32	−2338.8	4741.5	4947.2
Model 4	Multidimensional	Yes	28	−2302.7	4661.4	4841.4
Model 5	Bi-factor multidimensional	No	23	−2411.7	4869.4	5017.3
Model 6	Bi-factor multidimensional	Yes	22	−2328.0	4700.1	4841.4

Table 6. Estimates (Est.) and standard errors (SEs) of bi-factor explanatory longitudinal item response model

		Est.			SEs
Fixed effects		0.573			0.955
$\zeta_{\cdot\cdot}$ [Intercept]					
Persons					
ζ_{21} [Time 2 by Group]		−0.366			0.501
ζ_{31} [Time 3 by Group]		0.636			0.653
$\zeta_{2\cdot}$ [Time 2]		0.666*			0.138
$\zeta_{3\cdot}$ [Time 3]		2.775*			0.191
$\zeta_{\cdot 1}$ [Group]		−1.058*			0.191
Items					
β_1 [Operation]		−1.630*			0.881
β_2 [Measurement]		0.159			0.513
β_3 [Representation]		−0.584			0.682
Random effects					
Persons					
		Var.	Corr. Time 1	Time 2	
Σ_0 [Overall]	Time 1	0.440			
	Time 2	0.123	–		
	Time 3	1.016	–	–	
Σ_1 [Operation]	Time 1	0.389			
	Time 2	0.594	–		
	Time 3	0.110	–	–	
Σ_2 [Measurement]	Time 1	1.923			
	Time 2	1.323	–		
	Time 3	1.067	–	–	
Σ_3 [Representation]	Time 1	2.820			
	Time 2	0.001	–		
	Time 3	0.001	–	–	
Items					
σ_ϵ [Variance]		0.634			

Notes. *Significance based on p-value <.05 for fixed effects.
– Not modelled.

multidimensionality. Explanatory models (models 2 and 4) fit better than descriptive models (models 1 and 3). Model 4 shows the best fit to the data based on both AIC and BIC.

5. Bi-factor explanatory longitudinal item response models

Explanatory longitudinal item response models, explained in Section 3, provide dimension-specific change scores based on items in the same skill area, with the assumption that item responses are independent given θ_{jtd}. In the model, the covariation (or correlation) structure among specific dimensions within an item group, $\boldsymbol{\theta}_{jd} = [\theta_{j1d}, \ldots, \theta_{jtd}, \ldots, \theta_{jTd}]' \sim MN(\boldsymbol{\mu}_{d(T\times 1)}, \Sigma_{d(T\times T)})$, is modelled. Instead of having dimension-specific change scores, one may be interested in obtaining the *overall* change scores from *all* items in a test and dimension-specific change scores from items in an item group (or skill domain), with the assumption that there is a *general* dimension (or factor, θ_{jt0}) to be accounted for in item responses. In such a case, a bi-factor type structure can be imposed on items instead of allowing the covariation or correlation among specific dimensions (Gibbons & Hedeker, 1992).

In this section, we first present a bi-factor explanatory longitudinal item response model and then show how this can be presented as a special case of the generalized explanatory longitudinal item response model (equation (2)) with an empirical example.

A bi-factor explanatory longitudinal item response model is given by

$$\text{logit}[P(y_{g[j]d[i]t} = 1 \mid \theta_{jt0}, \theta_{jtd}, \epsilon_{d[i]})] = \zeta_{..} + (\zeta_{tg} + \zeta_{t.} + \zeta_{.g} + \theta_{jt0} + \theta_{jtd}) - (\beta_d + \epsilon_{d[i]}), \qquad (3)$$

where

- θ_{jt0} is a *general* random time parameter and
- θ_{jtd} is a dimension-specific random time parameter.

$\boldsymbol{\theta}_{j0}$ is assumed to follow a multivariate normal (*MN*) distribution, $\boldsymbol{\theta}_{j0} = [\theta_{j10}, \ldots, \theta_{jt0}, \ldots, \theta_{jT0}]' \sim MN(\boldsymbol{\mu}_{0(T\times 1)}, \Sigma_{0(T\times T)})$. $\boldsymbol{\theta}_{jd}$ is assumed to follow a multivariate normal (*MN*) distribution for each item group d, $\boldsymbol{\theta}_{jd} = [\theta_{j1d}, \ldots, \theta_{jtd}, \ldots, \theta_{jTd}]' \sim MN(\boldsymbol{\mu}_{d(T\times 1)}, \Sigma_{d(T\times T)})$. In order to identify the model, means ($\boldsymbol{\mu}_{\theta 0(T\times 1)}$ and $\boldsymbol{\mu}_{\theta d(T\times 1)}$) and covariances in $\Sigma_{0(T\times T)}$ and $\Sigma_{d(T\times T)}$ are set to $\mathbf{0}$.

5.1. Illustration of models using the FOC test data

A bi-factor explanatory longitudinal item response model can be framed as the generalized explanatory longitudinal item response model by specifying a different $Q_{d[i]t}$ for change. For a general dimension from all items, d is denoted by 0, yielding $Q_{0[i]t}$. As shown in Table 4, the Wiener simplex for the overall change in a general dimension is specified using $Q_{0[i]t}$ (such as $Q_{0[i]1}$, $Q_{0[i]2}$, and $Q_{0[i]3}$) in Table 4. This is also done for each dimension, $Q_{1[i]1}$, $Q_{1[i]2}$, and $Q_{1[i]3}$ for number operation items; $Q_{2[i]1}$, $Q_{2[i]2}$, and $Q_{2[i]3}$ for measurement items; and $Q_{3[i]1}$, $Q_{3[i]2}$, and $Q_{3[i]3}$ for representation items for times 1, 2, and 3, respectively.

For the purpose of comparison, both descriptive and explanatory bi-factor longitudinal item response models were considered:

- model 5, descriptive bi-factor longitudinal model for multidimensional tests (without fixed effects in equation (3));
- model 6, explanatory bi-factor longitudinal model for multidimensional tests (equation (3)).

The R lmer code for models 5 and 6 is shown in Appendix A. The data structure shown in Tables 3 and 4 is matched with the R lmer code in Appendix B.

Table 6 shows the parameter estimates of the bi-factor explanatory longitudinal item response model. Estimates of the fixed effects for items and persons from the model are similar to those from the explanatory longitudinal item response model (see Table 2). Instead of having correlations between the time dimensions for each skill domain, the variances of the general dimension and variances of dimension-specific dimensions are obtained in the bi-factor type model.

As shown in Table 5, an explanatory model (model 6) fits better than a descriptive model (model 5) based on AIC and BIC. Within descriptive models (models 1, 3, and 5), multidimensional models (models 3 and 5) fit better than a unidimensional model (model 1), and a bi-factor multidimensional model (model 5) does not fit better than a multidimensional model (model 3) based on AIC and BIC. Within explanatory models (models 2, 4, and 6), multidimensional models (models 4 and 6) fit better than a unidimensional model (model 2) based on AIC and BIC. Between the multidimensional models (models 4 and 6), a multidimensional model (model 4) fits better than a bi-factor multidimensional model (model 6) based on AIC, but the two models fit the data equally well based on BIC.

5.2. Score comparisons among models using the FOC test data

Since overall change estimates for unidimensional tests are often provided in practice, initial and change scores were compared across the models for unidimensional tests (models 1 and 2) and the models for multidimensional tests (models 3–6). The scores from the explanatory models (models 2, 4, and 6) can be considered as residual scores after the scores were explained by covariates (i.e., time points, a person group, and their interaction). For score comparison across the models, the following score calculation was considered:

- model 1 (Embretson descriptive model for unidimensional tests). Overall change scores based on all items are calculated for comparison purposes. It is likely that the change scores are not valid since the FOC test is multidimensional. These overall change scores can be compared with the overall change scores from the other models.
- model 2 (Embretson explanatory model for unidimensional tests). Similar to model 1, the overall change scores based on all items were obtained.
- model 3 (descriptive longitudinal model for multidimensional tests). This model calculates dimension-specific change scores based on items in the same skill area. In addition, the correlations between dimensions within a skill domain are estimated. Overall change scores are obtained by averaging the dimension-specific change scores.
- model 4 (explanatory longitudinal model for multidimensional tests). Similar to model 3, dimension-specific change scores based on items with the same skill area are obtained while estimating the correlations between dimensions within a domain. The overall change scores are obtained by averaging the dimension-specific change scores.
- model 5 (descriptive bi-factor longitudinal model for multidimensional tests). This model calculates both overall change and dimension-specific change scores. The overall change to be compared with other models is calculated by the sum of the overall change scores and the average of the three dimension-specific change scores at each time point.
- model 6 (explanatory bi-factor longitudinal model for multidimensional tests). Akin to model 5, the overall change to be compared with other models is calculated by the sum

Table 7. Comparisons of the overall initial values and changes across models

		Descriptive			Explanatory		
Time point		Overall (model 1)	3-dimension specific (model 3)	Overall and dimension specific specific (model 5)	Overall (model 2)	3-dimension specific (model 4)	Overall and dimension specific specific (model 6)
Time 1	Overall	1.000			1.000		
	3-dimension specific	0.731	1.000		0.753	1.000	
	Overall and dimension specific	0.832	0.933	1.000	0.825	0.941	1.000
Time 2	Overall	1.000			1.000		
	3-dimension specific	0.925	1.000		0.921	1.000	
	Overall and dimension specific	0.947	0.943	1.000	0.967	0.943	1.000
Time 3	Overall	1.000			1.000		
	3-dimension specific	0.779	1.000		0.750	1.000	
	Overall and dimension specific	0.913	0.725	1.000	0.933	0.683	1.000

of the overall change scores and the average of the three dimension-specific change scores at each time point.

To summarize, models 1 and 2 provide overall initial and change scores, models 3 and 4 provide three dimension-specific initial and change scores: with a correlation structure among dimensions and models 5 and 6 provide overall and dimension-specific change scores. The correlations in Table 7 depict the relationship between overall scores from models 1–6. The pattern of correlation is similar in the descriptive and explanatory models. As shown in Table 7, there is evidence that the overall change scores from models for unidimensional tests (models 1–2) and models for multidimensional tests (models 3–6) are not perfectly correlated, indicating that the change scores can be different when the multidimensionality of the test is ignored. In addition, the overall change scores are not perfectly correlated from models with a correlation structure among dimensions (models 3 and 4) and models with a bi-factor structure (models 5 and 6). This is because the correlation structure among dimensions within an item group (i.e., a skill domain) was ignored in calculating the overall change scores.

6. Summary and discussion

6.1. Summary

This paper has presented explanatory longitudinal item response models for multidimensional tests, which are new. There are two main extensions. The first involves the multidimensional extension of Embretson's (1991) model, specifically by specifying Embretson's (1991) model for each item group. The second extension involves using the explanatory approach of Embretson (1991) on both the person and item parameters of the item response models. The explanatory approach in longitudinal item response models is useful when one is interested in a multiple person group comparison of change and in the change score interpretation linked to item attributes. A bi-factor explanatory longitudinal item response model was also presented to obtain overall change scores and dimension-specific change scores. These models were represented within a generalized explanatory longitudinal item response model so one can see the similarities and differences among models and fit all these models using open source software, the R lmer function.

In the empirical application, we illustrated the practical uses of the models described in the paper. In the example analyses, the models were used to show the change score for each skill domain, the change score difference between LD students and non-LD students, and the change score interpretation based on relative complexity characterized by item difficulty across the different cognitive skills with a random residual across items for item difficulties. The empirical results showed that the overall change scores from all items may be different when the multidimensionality of a test is ignored, as illustrated in comparisons between the models for unidimensional tests (models 1 and 2) and the models for multidimensional tests (models 3–6). In a bi-factor explanatory longitudinal item response model, the overall change scores from all items and dimension-specific change scores were reported instead of having the covariation (or correlation) structure among time dimensions within a skill domain. We recommend the use of the bi-factor explanatory longitudinal item response model to calculate the overall change scores and provide more diagnostic change scores unless one is interested in investigating the covariation (or correlation) between time dimensions in an explanatory longitudinal item response model.

6.2. Discussion

6.2.1. Related models

Change is often measured using total scores as repeated measures, as in multilevel linear models (Bryk & Raudenbush, 1987; Pinheiro & Bates, 2000). Latent growth curve models (Rao, 1958; Tucker, 1958) and their extensions have been used to quantify change predominantly using structural equation models, historically fitted to a mean vector and a (co-)variance matrix of the total scores measured over time. Within these models, the individual differences in change are explained by the individual variation around an overall mean intercept (initial status) and a mean slope (change rate). These approaches start from the point at which we have already established the dimensionality of the latent construct at each time point, which then allows us to test the invariance of item parameters relating them to the constructs. To test measurement invariance and measure change simultaneously, we can use the extensions of these two approaches such as multilevel linear and nonlinear models (Pastor & Beretvas, 2006; Raudenbush, Johnson, & Sampson, 2003), a second-order growth model (Hancock, Kuo, & Lawrence, 2001; Sayer & Cumsille, 2001), and a unified latent curve and latent state-trait model (Tisak & Tisak, 2000). Using the multilevel linear and latent growth curve models, information on change for each person is often represented in terms of an initial status and the rate of change over time. Results from the longitudinal item response models provide the initial score and the change score at each time point for each person.

Fischer (1976) proposed a longitudinal item response model called the linear logistic model with relaxed assumptions (LLRA), which can be used with a multidimensional test to measure change in a construct over time under the assumption that each item corresponds to a dimension. Fischer's model is attractive since it does not require unidimensionality of the items or distributional assumptions about the population of persons. However, the LLRA has limitations. First, a person's change is explained by covariates such as a person group and time trend variables, but no individual differences in change (residuals after covariates are taken into account) are modelled. Second, parameter estimation in a LLRA is based on conditional maximum likelihood estimation (CMLE). CMLE uses the part of the data in an item for which change is observable (for example, $y_{t=1} = 0$ and $y_{t=2} = 1$ or $y_{t=1} = 1$ and $y_{t=2} = 0$ for two time points, where y is an item response and t is a time point), while ignoring remaining data showing no change. When there is the lack of change (i.e., $y_{t=1} = 0$ and $y_{t=2} = 0$ or $y_{t=1} = 1$ and $y_{t=2} = 1$ for two time points), the change parameter estimates diverge (see the Appendix in Formann & Spiel, 1989, for details). Thus, the estimation of both a person group effect and the trend effect becomes impossible (Fischer, 1989; Formann & Spiel, 1989). As a property of CMLE, if each item shows small change from one time point to another, the estimation method uses only a small portion of the data, so that the parameter estimates may become extremely inaccurate. This limitation of CMLE in estimating parameters in IRT is also noted by Tuerlinckx, Rijmen, Verbeke, and De Boeck (2006) and Verbeke, Spiessens, and Lesaffre (2001).

To prevent the second limitation of CMLE, Fischer (1989) extended the LLRA to a hybrid LLRA by adding different item difficulty levels at each time point, while measuring the same dimension. Therefore, the greater part of change may be compensated for by the difference in difficulty so that even small changes can be estimated with greater accuracy. However, this modification is only applicable to Rasch family models where the amount of change in a person dimension and item difficulty parameters is the same. In addition, change is confounded by the differences in item difficulty over time points. Since these

differences are the same for all persons, individual differences in change are not modelled.

In the generalized explanatory longitudinal item response model we have discussed, the limitations of the LLRA were overcome since the model presentation can be generalized to non-Rasch models and individual differences in change were modelled. The term 'generalized' in this study is used for two reasons: (1) the models described fall under the generalized linear and nonlinear mixed effect models (De Boeck & Wilson, 2004; McCulloch & Searle, 2001) or the generalized latent variable model (Skrondal & Rabe-Hesketh, 2004); and (2) different kinds of longitudinal item response models can be specified by changing only the indicators in the model. The motivation for presenting the generalized longitudinal item response model is in line with the motivation to formulate item response models as generalized linear and nonlinear mixed effect models (De Boeck & Wilson, 2004; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). First, it is helpful to understand the similarities and differences among different types of longitudinal item responses by applying the same framework to them. Second, software packages developed for generalized linear and nonlinear models can be used to fit the different types of longitudinal item response models. The R lmer function was chosen to fit the models in this paper. The R lmer function has been used for various item response models (see De Boeck *et al.*, 2011).

The explanatory models presented in this paper can be considered longitudinal extensions of explanatory item response models (De Boeck & Wilson, 2004) and multilevel item response models (e.g., Fox & Glas, 2001; Kamata, 2001; Maier, 2001, 2002; Rabe-Hesketh, Skrondal, & Pickles, 2004). In addition, they are longitudinal model extensions of linear logistic test models (Fischer, 1973), with the exception that a random residual for item difficulty parameters is allowed in this study. Further, it becomes an extension of a model discussed by Embretson (1995) by adding a random residual for item difficulty parameters.

6.2.2. Person group differences in longitudinal studies

Longitudinal data can be collected with person group information such as random assignment to different conditions in an experimental design (Bock, 1975) or different demographic groups in a longitudinal panel design (Hsiao, 2003).

Previous studies showed the importance of IRT scaling in estimating these group differences. Maxwell and Delaney (1985) demonstrated that a *t* test comparing groups on the observed test scores produces correct inferences at the construct level if and only if group variances on the construct are homogeneous or the difficulty level of the test equals the mean ability level of examinees. Embretson (1994, 1996) showed that observed total score approaches are problematic in that scaling artefacts produce significant differences in observed scores even when the latent true scores do not differ. Embretson (1996) also found that spurious interaction effects between groups and conditions (e.g., control vs. treatment) can be found in analysis of variance (ANOVA) designs using the total scores, and that test difficulty levels determine both the direction and magnitude of these biased interaction effects.

Longitudinal item response models offer some improvement over observed score approaches for modelling group differences. For example, Andersen (1980) showed how the means, variances, and correlation of two dimensions across two time points can be compared in the latent distributions without modelling individual differences in change. As a two-stage procedure, one can obtain change scores from longitudinal item response

models and then use ANOVA to investigate group differences in the change scores or the scores at each time point. However, as is frequently pointed out, the use of a two-stage procedure to investigate group differences may yield distorted mean estimates because measurement errors in the estimates are not incorporated appropriately. Person group differences were simultaneously modelled along with individual differences in the generalized explanatory longitudinal item response model.

6.2.3. Item attributes in longitudinal studies

There are two lines of research which match item attributes with change scores to interpret the change scores in terms of item attributes of interest. The first line of research is domain-referenced score reporting for change (Bock, Thissen, & Zimowski, 1997). The second line of research links change or learning to item attributes in terms of item groups such as specific processing mechanisms and knowledge structures (Embretson, 1995). In this study, the effects of item attributes were estimated taking unexplained variance in item difficulties into account and were used to interpret the change scores. When the item group information is misspecified, the change score interpretation can be distorted. Item information validation is required when using explanatory longitudinal item response models.

6.2.4. Measurement invariance testing

In the empirical illustration, DIF analyses were performed across person groups at each time point and then across time points based on all items in a test. When the test is assumed to be multidimensional, measurement invariance should be checked based on items in the same item groups for each dimension. However, when the number of item groups is small and the number of persons in a person group is small, it may not be feasible to use IRT DIF procedures to check measurement invariance because the parameter estimates are not precise or stable. IRT DIF detection in the empirical illustration assumed that the person parameter estimates to be conditioned upon are unidimensional. DIF approaches based on total scores were implemented to test DIF across person groups at each time point because the number of persons was small. DIF analysis for multidimensional tests is an area of ongoing research. In future research it would be useful to investigate measurement invariance over time when the test is multidimensional.

The same set of items over multiple time points was used in the applications. Memory effects, response consistency effects, and practice effects are all potential confounds that exist when dealing with repeated measures. This could possibly result in a violation of the local independence assumption within a skill domain in measuring change. However, the FOC test items were performance tasks with multiple steps embedded in each. For example, some items required students to interpret the schematic plans of a building project, determine the most economical use of wood and other materials, and compute the total cost. Given that students were not shown test results or correct/incorrect answers, it is unlikely that memory effects affected overall test performance.

6.2.5. Parameter estimation

The generalized longitudinal item response model (equation (2)) is a special case of generalized linear and nonlinear mixed models. In explanatory and bi-factor explanatory

longitudinal item response models (equations (1) and (3)), the random effects, $\theta_{jtd}/\theta_{jt0}$ and $\epsilon_{d[i]}$, are *crossed*, not nested as they are in generalized longitudinal item response models. Unfortunately, maximum likelihood estimation of models for categorical data is challenging. This is because the marginal likelihood does not have a closed form, thus requiring estimation with numerical or Monte Carlo integration. It is even more challenging for the model with crossed random effects. If the random effects are nested, the integrals are also nested (e.g., Rabe-Hesketh, Skrondal, & Pickles, 2005), keeping the computational burden low, but for crossed random effects, high-dimensional integrals need to be evaluated. There are three distinct solutions to this problem for generalized linear and nonlinear mixed models: (1) approximation of the integral with numerical integration techniques such as (adaptive) Gauss–Hermite quadrature and Monte Carlo integration; (2) approximation for the integrand; and (3) simulation-based methods such as Markov chain Monte Carlo (MCMC). See Cho and Rabe-Hesketh (2011), Cho, Partchev, and De Boeck (2012), Rabe-Hesketh *et al.* (2004), and Rijmen *et al.* (2003) for an overview of these three estimation methods.

The R lmer function used in this study is based on the Laplace approximation for the integrand. The R lmer function is a flexible software application for various Rasch family item response models (see De Boeck *et al.*, 2011). The R lmer function is computationally efficient for high-dimensional item response models compared to an approximation for the integral and simulation-based methods such as MCMC (Cho *et al.*, 2012). The empirical data used in this study had a relatively small number of items and a small sample size. The computation time with the R lmer function ranged from 3 min to 12 min for all models considered in this study on a computer equipped with a 1.20 GHz processor with 2.93 GB of RAM. Given the high dimensionality of the models presented, the computation time will increase rapidly for large sample sizes and a large number of items. For example, it took 42 min for simulated data with 1,000 persons, 40 items, and three time points on the same computer.

Laplace approximation can perform poorly for a dichotomous response with small cluster sizes (e.g., the number of items), with a downward bias in the estimated variance components (Cho & Rabe-Hesketh, 2011; Goldstein & Rasbash, 1996; Raudenbush, Yang, & Yosef, 2000; Rodriguez & Goldman, 1995). The downward bias in variance components can lead to overestimated correlations (Cho *et al.*, 2012). These findings may explain the perfect correlation coefficients (-1 and 1 for the operation skill in Table 2) for the number operation skill because the number of items within a skill domain is small in our empirical data.

The standard errors associated with initial ability and change estimates can be obtained for each individual in the IRT framework. At the test level, an average reliability across individuals can also be obtained. In the longitudinal model context, it would be informative to report the average reliability at each time point. Unfortunately, there are limited models where the standard errors can be obtained using the extraction function se.ranef in R after the models are fitted using the R lmer function. The standard errors cannot be obtained for the explanatory models in the current application using the seranef function in R.

6.2.6. Model selection

The AIC and BIC were used to compare alternative models in the empirical study. The appropriate use of information criteria in latent variable models is an ongoing area of research (e.g., Vaida & Blanchard, 2005). Further study is required to investigate the

performance of AIC and BIC when compared with models which differ with respect to the use of explanatory variables and a different dimensionality.

6.2.7. Extensions of the models

Only Rasch versions of the models described in this paper were considered, although they can be expanded to more complex models. Small sample sizes limit feasibility for estimating item discrimination parameters in the illustrative study. Two-parameter extensions can be made easily by replacing the 1 in $Q_{d[i]t}$ with an item discrimination parameter α_i.

The explanatory model formulation in this study was based on the logistic item response model formulation followed by De Boeck and Wilson (2004). When the models are presented with normal ogive item response models, they can be easily reformulated in terms of categorical factor analysis within the structural equation modeling framework by adding a measurement model for the items (Takane & de Leeuw, 1987). A limited information estimation method was historically used in structural equation modelling. As far as we are aware, the limited estimation methods for crossed random effects are not currently feasible in commercial software. A new version of Mplus will in due course be available for models with crossed random effects using hierarchical Bayesian analysis (Asparouhov & Muthén, 2012).

Acknowledgements

We are grateful to Dr. Brian Bottge (University of Kentucky) for making the data available for the application. We thank Dr. Paul De Boeck (Ohio State University) for his valuable comments. We also thank two anonymous reviewers and associate editor Dr. Herbert Hoijtink for their insightful comments and suggestions.

References

Akaike, A. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. doi:10.1109/TAC.1974.1100705

Andersen, E. B. (1980). Comparing latent distributions. *Psychometrika*, *45*, 121–134. doi:10.1007/BF02293602

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16. doi:10.1007/BF02294143

Asparouhov, T., & Muthén, B. (2012). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models using s4 classes. Retrieved from: http://cran.r-project.org/web/packages/lme4/index.html

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores, *Journal of Educational Measurement*, *34*, 197–211. doi:10.1111/j.1745-3984.1997.tb00515.x

Bottge, B. A., Heinrichs, M., Chan, S. Y., Mehta, Z. D., & Watson, E. (2003). Effects of video-based and applied problems on the procedural math skills of average-and low-achieving adolescents. *Journal of Special Education Technology*, *18*, 5–22. Retrieved from: http://www.tamcec.org/jset/

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*, 147–158. doi:10.1037/0033-2909.101.1.147

Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics & Data Analysis*, *55*, 12–25. doi:10.1016/j.csda.2010.04.015

Cho, S.-J., Bottge, B., Cohen, A. S., & Kim, S.-H. (2011). Detecting cognitive change in the math skills of low-achieving adolescents. *Journal of Special Education*, *45*, 67–76. doi:10.1177/0022466909351579

Cho, S.-J., Partchev, I., & De Boeck, P. (2012). Parameter estimation of multiple item profiles models. *British Journal of Mathematical and Statistical Psychology*, *65*, 438–466.

De Boeck, P. (2011). Sequential order based GLMMs for random person and item effects. Paper presented at the International Meeting of the Psychometric Society, Hong Kong.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1–28. Retrieved from: http://www.jstatsoft.org/

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197. doi:10.1037/0033-2909.93.1.179

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515. doi:10.1007/BF02294487

Embretson, S. E. (1994). Comparing changes between groups: Some perplexities arising from psychometrics. In D. Laveault, B. D. Zumbo, M. E. Gessaroli & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 213–248). Ottawa: Edumetrics Research Group, University of Ottawa.

Embretson, S. E. (1995). A measurement model for linking individual learning to process and knowledge. *Journal of Educational Measurement*, *32*, 277–294. doi:10.1111/j.1745-3984.1995.tb00467.x

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*, 201–212. doi:10.1177/01466216960200302

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97–110). New York: Wiley.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika*, *54*, 599–624. doi:10.1007/BF02296399

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Formann, A. K., & Spiel, C. (1989). Measuring change by means of a hybrid variant of the linear logistic model with relaxed assumptions. *Applied Psychological Measurement*, *13*, 91–103. doi:10.1177/014662168901300109

Fox, J.-P., & Glas, C. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271–288. doi:10.1007/BF02294839

Gibbons R. D., & Hedeker D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436. doi:10.1007/BF02295430

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, *159*, 505–513. doi:10.2307/2983328

Hancock, G. R., Kuo, W., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 470–489. doi:10.1207/S15328007SEM0803_7

Hsiao, C. (2003). *Analysis of panel data*. Cambridge: Cambridge University Press.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). New York: Springer.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79–93. doi:10.1111/j.1745-3984.2001.tb01117.x

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, *26*, 307–330. doi:10.3102/10769986026003307

Maier, K. S. (2002). Modeling incomplete scaled questionnaire data with a partial credit hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, *27*, 271–289. doi:10.3102/10769986027003271

Maxwell, S. E., & Delaney, H. D. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin*, *97*, 85–93. doi:10.1037/0033-2909.97.1.85

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear and mixed models*. New York: Wiley.

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement*, *30*, 100–120. doi:10.1177/0146621605279761

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from: http: //www.R-project.org/

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190. doi:10.1007/BF02295939

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*, 301–323. doi:10.1016/j.jeconom.2004.08.017

Rao, C. R. (1958). Some statistical methods for the comparison of growth curves. *Biometrics*, *14*, 1–17. doi:10.2307/2527726

Raudenbush, S. W., Yang, M., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, *9*, 141–157. Retrieved from: http://www.jstor.org/stable/1390617

Raudenbush, S. W., Johnson, C., & Sampson, R. J. (2003). A multivariate, multilevel Rasch model for self-reported criminal behavior. *Sociological Methodology*, *33*, 169–211. doi:10.1111/j.0081-1750.2003.t01-1-00130.x

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205. doi:10.1037/1082-989X.8.2.185

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, *158*, 73–89. doi:10.2307/2983404

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 179–200). Washington, D. C.: American Psychological Association.

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, *16*, 143–152. doi:10.1111/j.1745-3984.1979.tb00095.x

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 416–464. doi:10.1214/aos/1176344136

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408. doi:10.1007/BF02294363

te Marvelde, J. M., Glas, G. A. W., Landeghem, G. V., & Van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement*, *66*, 5–34. doi:10.1177/0013164405282490

Tisak, J., & Tisak, M. S. (2000). Permanency and ephemerality of psychological measures with application to organizational commitment. *Psychological Methods*, *5*, 175–198. doi:10.1037/1082-989X.5.2.175

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika*, *23*, 19–23. doi:10.1007/BF02288975

Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, *59*, 225–255. doi:10.1348/000711005X79857

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed effects models. *Biometrika*, *92*, 351–370. doi:10.1093/biomet/92.2.351

Verbeke, G., Spiessens, B., & Lesaffre, E. (2001). Conditional linear mixed models. *American Statistician*, *55*, 25–34. doi:10.1198/000313001300339905

Received 5 November 2011; revised version received 12 May 2012

Supporting Information

The following supporting information may be found in the online edition of the article:

Table S1. Estimates and standard errors (SE) of model 1 using the simulation data.

Table S2. Estimates and standard errors (SE) of model 2 using the simulation data.

Table S3. Estimates and standard errors (SE) of model 3 using the simulation data.

Table S4. Estimates and standard errors (SE) of model 4 using the simulation data.

Table S5. Estimates and standard errors (SE) of model 5 using the simulation data.

Table S6. Estimates and standard errors (SE) of model 6 using the simulation data.

Table S7. Model fit comparisons using the simulation data.

Table S8. Model fit comparisons using the empirical data (Table 5 of the paper).

The simulated data similar to the empirical data.

R lmer script for models 1-6.

Appendix A: The lmer code for the empirical application

```
data <- read.table("C:/data.dat",header = T,fill = T)
    # Call libraries
    library(lme4)
    data$ITEM <- as.factor(data$item)
    data$TIME <- as.factor(data$time)
    data$GROUP <- as.factor(data$GROUP)
    # Model 1: Embretson descriptive model for unidimensional tests
    M1 <- lmer(y ~ -1 + ITEM + (Q1 + Q2 + Q3 -1 | person), data, binomial("logit"))
    # Model 2: Embretson descriptive model for unidimensional tests
    M2 <- lmer(y ~  1 + TIME*GROUP + operation + measurement + representation
+ (1 | item) + (Q1 + Q2 + Q3 -1 | person), data, binomial("logit"))
    # Creating Q_{d[i]t} for models 3, 4, 5, and 6
    data$operationQ1 <- (data$operation)*(data$Q1)
    data$operationQ2 <- (data$operation)*(data$Q2)
    data$operationQ3 <- (data$operation)*(data$Q3)
```

```
data$measurementQ1 <- (data$measurement)*(data$Q1)
data$measurementQ2 <- (data$measurement)*(data$Q2)
data$measurementQ3 <- (data$measurement)*(data$Q3)
data$representationQ1 <- (data$representation)*(data$Q1)
data$representationQ2 <- (data$representation)*(data$Q2)
data$representationQ3 <- (data$representation)*(data$Q3)
# Model 3: Descriptive longitudinal models for multidimensional tests
M3 <- lmer(y ~ -1 + ITEM + (operationQ1 + operationQ2 + operationQ3 -1 | person) +
(measurementQ1 + measurementQ2 + measurementQ3 -1 | person) + (representationQ1
+ representationQ2 + representationQ3 -1 | person), data, binomial("logit"))
# Model 4: Explanatory longitudinal models for multidimensional tests
M4 <- lmer(y ~ 1 + TIME*GROUP + operation + measurement + representation
+ (1 | item) + (operationQ1 + operationQ2 + operationQ3 -1 | person) + (measurementQ1
+ measurementQ2 + measurementQ3 -1 | person) + (representationQ1 + representa-
tionQ2 + representationQ3 -1 | person), data, binomial("logit"))
# Model 5: Descriptive bi-factor longitudinal models for multidimensional tests
M5 <- lmer(y ~ -1 + ITEM + (Q1 -1 | person) + (Q2 -1 | person) + (Q3 - 1 | person) + (oper-
ationQ1 -1 | person) + (operationQ2 -1 | person) + ( operationQ3 -1 | person) + (measur-
ementQ1 -1 | person) + (measurementQ2 -1 | person) + (measurementQ3 -1 | person) +
(representationQ1 -1 | person) + (representationQ2 -1 | person) + (representationQ3
-1 | person), data, binomial("logit"))
# Model 6: Explanatory bi-factor longitudinal models for multidimensional tests
M6 <- lmer(y ~ 1 + TIME*GROUP + operation + measurement + representation +
(1 | item) + (Q1 -1 | person) + (Q2 -1 | person) + (Q3 -1 | person) + (operationQ1
-1 | person) + (operationQ2 -1 | person) + ( operationQ3 -1 | person) + (measurementQ1
-1 | person) + (measurementQ2 -1 | person) + (measurementQ3 -1 | person) +
(representationQ1 -1 | person) + (representationQ2 -1 | person) + (representationQ3
-1 | person), data, binomial("logit"))
# Model comparisons
anova(M1,M2,M3,M4,M5,M6)
```

Appendix B: The data/indicators of the models and lmer codes

Model		Data	lmer syntax
Model 1	Table 3	$y_{g[j]d[i]t}$	y
		$X_{1[1]}, X_{1[2]}, \ldots, X_{1[13]}, X_{1[14]}$	-1 + ITEM
		$Q_{1[i]1}$	Q1
		$Q_{1[i]2}$	Q2
		$Q_{1[i]3}$	Q3
		j	person
Model 2	Table 3	$y_{g[j]d[i]t}$	y
		t	data$TIME <- as.factor(data$time)
		$Z_{0j}X_{0i}$	1
		$Z_{1[j]}$	data$GROUP <- as.factor(data$GROUP)
		$X_{1[i]}$	operation

Continued

Appendix B. (*Continued*)

Model		Data	lmer syntax
		$X_{2[i]}$	measurement
		$X_{3[i]}$	representation
		$Q_{1[i]1}$	Q1
		$Q_{1[i]2}$	Q2
		$Q_{1[i]3}$	Q3
		i	item
		j	person
Model 3	Table 4	$y_{g[j]d[i]t}$	y
		$X_{1[1]}, X_{1[2]}, \ldots, X_{1[13]}, X_{1[14]}$	-1 + ITEM
		$Q_{1[i]1}$	operationQ1
		$Q_{1[i]2}$	operationQ2
		$Q_{1[i]3}$	operationQ3
		$Q_{2[i]1}$	measurementQ1
		$Q_{2[i]2}$	measurementQ2
		$Q_{2[i]3}$	measurementQ3
		$Q_{3[i]1}$	representationQ1
		$Q_{3[i]2}$	representationQ2
		$Q_{3[i]3}$	representationQ3
		j	person
Model 4	Table 4	$y_{g[j]d[i]t}$	y
		t	data$TIME <- as.factor(data$time)
		$Z_{0j}X_{0i}$	1
		$Z_{1[j]}$	data$GROUP <- as.factor(data$GROUP)
		$X_{1[i]}$	operation
		$X_{2[i]}$	measurement
		$X_{3[i]}$	representation
		$Q_{1[i]1}$	operationQ1
		$Q_{1[i]2}$	operationQ2
		$Q_{1[i]3}$	operationQ3
		$Q_{2[i]1}$	measurementQ1
		$Q_{2[i]2}$	measurementQ2
		$Q_{2[i]3}$	measurementQ3
		$Q_{3[i]1}$	representationQ1
		$Q_{3[i]2}$	representationQ2
		$Q_{3[i]3}$	representationQ3
		i	item
		j	person
Model 5	Table 4	$y_{g[j]d[i]t}$	y
		$X_{1[1]}, X_{1[2]}, \ldots, X_{1[13]}, X_{1[14]}$	-1 + ITEM
		$Q_{0[i]1}$	Q1
		$Q_{0[i]2}$	Q2
		$Q_{0[i]3}$	Q3
		$Q_{1[i]1}$	operationQ1
		$Q_{1[i]2}$	operationQ2
		$Q_{1[i]3}$	operationQ3
		$Q_{2[i]1}$	measurementQ1
		$Q_{2[i]2}$	measurementQ2
		$Q_{2[i]3}$	measurementQ3

Continued

Appendix B. (*Continued*)

Model		Data	lmer syntax
		$Q_{3[i]1}$	representationQ1
		$Q_{3[i]2}$	representationQ2
		$Q_{3[i]3}$	representationQ3
		j	person
Model 6	Table 4	$y_{g[j]d[i]t}$	y
		t	data$TIME <- as.factor(data$time)
		$Z_{0j}X_{0i}$	1
		$Z_{1[j]}$	data$GROUP <- as.factor(data$GROUP)
		$X_{1[i]}$	operation
		$X_{2[i]}$	measurement
		$X_{3[i]}$	representation
		$Q_{0[i]1}$	Q1
		$Q_{0[i]2}$	Q2
		$Q_{0[i]3}$	Q3
		$Q_{1[i]1}$	operationQ1
		$Q_{1[i]2}$	operationQ2
		$Q_{1[i]3}$	operationQ3
		$Q_{2[i]1}$	measurementQ1
		$Q_{2[i]2}$	measurementQ2
		$Q_{2[i]3}$	measurementQ3
		$Q_{3[i]1}$	representationQ1
		$Q_{3[i]2}$	representationQ2
		$Q_{3[i]3}$	representationQ3
		i	item
		j	person