This article was downloaded by: [KU Leuven University Library]
On: 10 February 2015, At: 11:07
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



The Journal of Experimental Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vjxe20

Modeling Growth in Electronic Learning Environments Using a Longitudinal Random Item Response Model

Damazo T. Kadengye[a], Eva Ceulemans[a] & Wim Van Den Noortgate[a]
[a] University of Leuven, Kortrijk, Belgium
Published online: 09 Jun 2014.

CrossMark
Click for updates

To cite this article: Damazo T. Kadengye, Eva Ceulemans & Wim Van Den Noortgate (2015) Modeling Growth in Electronic Learning Environments Using a Longitudinal Random Item Response Model, The Journal of Experimental Education, 83:2, 175-202, DOI: 10.1080/00220973.2014.907226

To link to this article: http://dx.doi.org/10.1080/00220973.2014.907226

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms &

Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

THE JOURNAL OF EXPERIMENTAL EDUCATION, *83*(2), 175–202, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 0022-0973 print/1940-0683 online
DOI: 10.1080/00220973.2014.907226

Routledge
Taylor & Francis Group

Modeling Growth in Electronic Learning Environments Using a Longitudinal Random Item Response Model

Damazo T. Kadengye, Eva Ceulemans, and Wim Van Den Noortgate
University of Leuven, Kortrijk, Belgium

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

In educational environments, monitoring persons' progress over time may help teachers to evaluate the effectiveness of their teaching procedures. Electronic learning environments are increasingly being used as part of formal education and resulting datasets can be used to understand and to improve the environment. This study presents longitudinal models based on the item response theory (IRT) for measuring persons' ability within and between study sessions in data from web-based learning environments. Two empirical examples are used to illustrate the presented models. Results show that by incorporating time spent within- and between-study sessions into an IRT model; one is able to track changes in ability of a population of persons or for groups of persons at any time of the learning process.

Keywords *item response theory, electronic learning, growth modeling*

IN EDUCATIONAL SETTINGS, students can build and improve their knowledge by being taught in traditional classroom environments, in electronic learning (e-learning) environments, or with a combination of both. A specific kind of environment is an item-based e-learning environment in which students (henceforth referred to as persons) can freely engage exercise items online, and answers are centrally logged. The FraNel project (Desmet, Paulussen, & Wylin, 2006) and the Maths Garden (Klinkenberg, Straatemeier, & Van der Maas, 2011) are good examples of such e-learning environments. A person gets instant feedback on the engaged item (e.g., about the correctness of the answer or a more elaborate feedback) with the purpose that learning takes place. This is similar to a scenario in which a teacher poses a question to persons in traditional classroom environments and then gives instant feedback to the given responses. The resulting data can be logged (i.e., responses to items or other data can be centrally recorded).

Tracking and logging data can be very informative. In e-learning environments, as in other learning environments, a person's learning progress is influenced by properties of the learning environment, properties of the person, and the interaction between both kinds of properties (e.g., Nam & Smith-Jackson, 2007; Welsh, Wanberg, Brown, & Simmering, 2003), and tracking and logging data can give information about these effects. Statistical analyses of such data can help us get a view on, for example, the ability level and preferences of persons, and on the difficulty

Address correspondence to Damazo T. Kadengye, Faculty of Psychology and Educational Sciences, University of Leuven (Kulak), Etienne Sabbelaan 53, 8500 Kortrijk, Belgium. E-mail: dkadengye@gmail.com

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/vjxe.

and psychometric quality of the engaged items. One way to achieve this is by using item response theory (IRT; Van der Linden & Hambleton, 1997) models. The use of IRT is an example of a data-centric modeling approach that estimates the ability of the persons on the basis of their performance (Chen, Lee, & Chen, 2005; Johns, Mahadevan, & Woolf, 2006). Furthermore, IRT offers new perspectives to e-learning environments, for example, automatically adapting the learning environment to the learner by rendering items of a difficulty that matches the ability of the person (Klinkenberg et al., 2011; Wauters, Desmet, & Van den Noortgate, 2010).

Measuring a person's ability at any time of the learning process may be compared with the concept of diagnostic teaching (Guskey, 2003) where teachers evaluate a person's learning on the spot or collect data at different points in time and compare progress over units of instruction. Within a diagnostic teaching perspective, assessment and instruction are interacting and continuous processes, with assessment providing feedback to the teacher on the efficacy of prior instruction, and new instruction building on the learning that persons demonstrate over time. Monitoring persons' progress during a study session may help teachers to continually evaluate the effectiveness of their teaching or training procedures and make more informed instructional decisions (Safer & Fleischman, 2005). This is important because some training procedures may actually result in slow or no improvement in the ability for which the training was implemented, or worse still, may cause a decrease in the ability level it was intended to improve (Miceli, Settanni, & Vidotto, 2008). If we want the e-learning environment to render exercise items that are adapted to the ability level of, for example, groups of persons, it is important to build statistical models such that at any moment of the learning process, that is within and between study periods, we can estimate reliably the average ability levels of these groups of persons.

In the IRT framework, longitudinal item response models (Andersen, 1985; Andrade & Tavares, 2005; Embretson, 1991; Meiser, Stern, & Langeheine, 1999; Miceli et al., 2008) have been proposed to model individual differences in growth over measurement occasions or time. One common situation is that at some specific measurement occasions, item sets are rendered to all persons. This may, for example, be in typical pretest and posttest studies where each person receives the same sets of items at different measurement occasions (e.g., see Andersen, 1985). Another example is in some educational settings where students are asked to take different sets of comparable tests at different predefined time intervals of an academic year (e.g., see Andrade & Tavares, 2005); or in studies where a group of students is followed from one grade to another and at the end of each academic year, a test is administered to these students. In the later case, at every measurement occasion, each person receives alternate sets of items that are not necessarily comparable in difficulty but may have some common items across the occasions (Embretson, 1991; Roberts & Ma, 2006). In these examples, it is assumed that some kind of intervention is introduced between blocks or between sets of exercise items that are presented to persons (Verguts & De Boeck, 2000; Verhelst & Glas, 1995), such that persons' average ability increases, or a group of items' difficulty decreases (Embretson, 1991; Fischer, 1995; te Marvelde, Glas, Landeghem & Damme, 2006).

In the cited examples so far and for the IRT models used therein, growth in individuals' ability is measured as having occurred between, but not within the measurement occasions and are thus not the focus of the present study. This is because for e-learning environments, making items is not primarily meant to measure the ability at that moment, but rather to enhance learning, for example by giving elaborated feedback after each response given by the user. The feedback and hints provided to persons can be looked at as a form of intervention intended to increase persons'

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

ability (Verguts & De Boeck, 2000). Therefore, learning can be achieved not only between measurement occasions (c.q., the time between two successive study sessions of the same course unit), but also during the measurement occasions (c.q., while making exercises within a lesson or study session). If we would ignore learning during a session and assume fixed study sessions, existing longitudinal IRT models, for example that of Andersen (1985) or of Embretson (1991) can be employed, estimating the average ability of persons for each study session. However, as described in the "Data Structure" section, the datasets considered in the present article differ in such a way that, the (number of) study sessions can depend on the person as a result of user freedom, and that growth in ability is expected both within and between study sessions. Furthermore, item sets are not (strictly) the same across all persons for each study session. This is because items are answered in various positions of the learning process by different persons, allowing for disentangling of item difficulties and time trends.

In this study, therefore, we present and develop longitudinal IRT models for modelling growth in persons' ability within and between study sessions for tracking and logging data from e-learning environments. First, a model with separate slopes for each study session is presented and illustrated with an empirical example. Second, a model with identical slopes for each study session (simplifying the first model), but also including between-session slopes is presented and illustrated with two empirical examples from web-based learning environments. The aim of the presented models is to allow e-learning instructors and managers to efficiently monitor persons' progress during the learning process and possibly personalize individuals' learning paths. Explanatory extensions of these models are also presented and illustrated with the mentioned examples. In the remaining sections of this article, we describe a general structure of a tracking and logging dataset. Next, the proposed models are presented and illustrated with empirical datasets. Last, we conclude with a discussion and a note on possible areas of improvement for future research.

Data Structure

Suppose P persons are participating in an item-based e-learning program. Persons log in to the learning environment and freely engage exercises, before logging out. This is referred to as a study session. We will refer to the period or space between two consecutive study sessions of a person as the between-session time. For example, if for a specific person the time between two consecutive responses is at least one day, we could consider in some situations the item responses as stemming from two separate study sessions (cf. Cepeda, Vul, Rohrer, Wixted, & Pashler; 2008; Vlach & Sandhofer; 2012).

Let k ($k = 1, 2, \ldots, K_p$) denote a study session, with K_p the total number of sessions for person p($p = 1, 2, ..., P$). Also, let i ($i = 1, 2, \ldots, N$) be an indicator for the item with N the totality of all items in the learning environment, and n_{pk}be the number of exercise items engaged by person p during a study session k. Take a hypothetical example in Table 1. In Table 1, we assume that a person ($p = 1$) engaged items in three study sessions ($k = 1, 2, 3$), and had two between-session periods, one of 24 hr, and another of 30 hr. He engaged n_{11}, n_{12} and n_{13} items in sessions 1, 2 and 3 respectively. Each item was engaged at a real-time point *time* (in hours) as shown in the third column of Table 1. From this time variable, one can obtain session-specific measurement time points denoted *stime*. Note that the values in Table 1 are arbitrary, simply used here for clarification purposes.

TABLE 1
Hypothetical Example of a Person's Properties and Session-Specific Time Indicators

Person	*Session*	*Time (hrs)*	*Item*	*Score*	stime	wtime	btime
1	1	0.00	1	1	0.00	0.00	0
1	1	0.10	2	0	0.10	0.10	0
1	1	0.20	3	1	0.20	0.20	0
1	1	.	.	.	.	.	.
1	1	.	.	.	.	.	.
1	1	t1	n_{11}	.	t1	t1	0
1	2	t1 + 24 + 0.00	$n_{11} + 1$	0	0.0	t1 + 0.00	24
1	2	t1 + 24 + 0.01	$n_{11} + 2$	1	0.01	t1 + 0.01	24
1	2	t1 + 24 + 0.15	$n_{11} + 3$	1	0.15	t1 + 0.15	24
1	2	.	.	.	.	.	.
1	2	.	.	.	.	.	.
1	2	t1 + t2 + 24	$n_{11} + n_{12}$	1	t2	t1 + t2	24
1	3	t1 + t2 + 54 + 0.00	$n_{11} + n_{12} + 1$	0	0.00	t1 + t2 +0.00	54
1	3	t1 + t2 + 54 + 0.02	$n_{11} + n_{12} + 2$	1	0.02	t1 + t2 + 0.02	54
1	3	t1 + t2 + 54 + 0.12	$n_{11} + n_{12} + 3$	1	0.12	t1 + t2 +0.12	54
1	3	.	.	.	.	.	.
1	3	.	.	.	.	.	.
1	3	t1 + t2 + t3 + 54	$n_{11} + n_{12} + n_{13}$	1	t3	t1 + t2 + t3	54

Note. *wtime* is the within-study sessions time, and *btime* is the between-study sessions time.

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

A Longitudinal IRT Model

One of the commonly used IRT models for data with binary item response scores is the Rasch model (Rasch, 1993). The strength of the Rasch model as the IRT model of choice is that, unlike most common IRT models, it can be reformulated within the generalized (non)linear mixed modeling framework, and therefore can be easily extended by using observed person and/or item properties with fixed or random effects (cf. De Boeck & Wilson, 2004; De Boeck et al., 2011). Rasch models are specific generalized linear mixed models because the logit or probit transformations of the expected values of a binary variable (i.e., probabilities) are linearly related to predictor variables, with one or more regression weights (or the latent traits) in the resulting linear model as random variables. The generalized linear mixed modeling framework is appealing in this context, because we are interested in further modeling within and between study session evolutions of person's abilities, by including time variables as covariates.

In the Rasch model, the probability π_{pi} of a correct response by person p to item i is regarded as a function of a person and an item parameter, θ_p and β_i, interpreted as the person's ability and item difficulty, respectively:

$$\begin{aligned} \text{logit}\left(\pi_{pi}|\theta_p, \beta_i\right) &= \ln\left(\pi_{pi}/1 - \pi_{pi}\right) \\ &= \theta_p - \beta_i \text{and} Y_{pi} \sim \text{binomial}(1, \pi_{pi}). \end{aligned} \quad (1)$$

where Y_{pi} denotes a binary response (1 = correct, 0 = wrong) of person $p$to item $i (i = 1, 2, ..., N)$. In IRT modeling one often assumes that the persons are a random sample from a population in which the person abilities are normally distributed with $\theta_p \sim N(0, \sigma_\theta^2)$, and the items have a

fixed difficulty. In Equation 1, the person parameter θ_p is a constant, but in our application, we expect an individual's ability to change over time, for example, as a result of learning. As such, the probability of a correct response by person p to item i can be thought of as a function of a person's ability at the time person p answers item i (θ_{pi}), and an item difficulty parameter, β_i, such that Equation 1 is reformulated as follows:

$$\text{logit}\left(\pi_{pi}|\theta_{pi}, \beta_i\right) = \theta_{pi} - \beta_i. \tag{2}$$

Note that Equation 2 assumes that an item i was solved at most once by person p. If this is not the case, an additional index can be used to indicate the time point on which the item was solved. The item difficulty parameter β_i in Equation 2 remains the same over time but we allow for changes in a person's ability parameter, θ_{pi}, at any time within any given study session. In case interest is in making inferences on the population of items rather than on individual items, they too can be regarded as random variables (cf. De Boeck, 2008; De Boeck & Wilson, 2004; Van den Noortgate, De Boeck & Meulders; 2003), such that the item difficulties are assumed to be random and normally distributed over items, $\beta_i \sim N(0, \sigma_\beta^2)$. Modeling items as random is a particularly appealing approach in situations involving large samples of items, because it results in a parsimonious model. By regarding the items as random, there is still room in the model to include item characteristics as predictors in the explanatory models, whereas including item predictors would make the model overidentified in case item parameters are considered as fixed (Van den Noortgate et al., 2003).

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

Measuring Growth Within Study Sessions

A person's ability θ_{pi} can be thought of as varying, possibly as linearly changing over time with session-specific slopes (rates of change in ability) that can vary over sessions and over individuals. In this regard, the actual ability of a person at any time can be regarded as a combination of a person's initial ability before a particular study session and the changes on this ability realized over time within this session, probably as a result of learning. Therefore, the probability of a correct response by person p to item i can be reformulated from Equation 2 as follows:

$$\text{logit}\left(\pi_{pi}\right) = \sum_{k=1}^{K}(\alpha_{0k} + \omega_{0pk})D_{pik} + \sum_{k=1}^{K}\left(\alpha_{1k} + \omega_{1pk}\right) stime_{pi} D_{pik} + \upsilon_i, \tag{3}$$

where

D_{pik} is a dummy variable equal to 1 if item i was solved by person p in session k, zero otherwise. Therefore, α_{0k} is the expected logit of the success probability for an average person on an average item at the start of session k i.e. at time zero; ω_{0pk} is the deviation of person p from the overall intercept α_{0k} for session k; α_{1k} is the overall slope for session k; ω_{1pk} is the deviation of person p within session k from the overall slope α_{1k}; and υ_i are random item effects with $\upsilon_i \sim N(0, \sigma_\upsilon^2)$.

The person random intercepts (ω_{0pk}) and random slopes (ω_{1pk}) may be assumed to be multivariate normally distributed with $N(0, \Sigma)$ where Σ is the variance-covariance matrix. From Equation 3, it can be deduced that the value $\theta_{pi} = \alpha_{0k} + \omega_{0pk} + (\alpha_{1k} + \omega_{1pk})stime_{pi}$ corresponds to the ability of person p during study session k at the time of responding to item i.

Similarly, the random item effects, υ_i, can be interpreted as easiness levels of items, $-\beta_i$ of Equation 2. In this model, items need not be the same for all individuals in a specific study session. However, if a specific session k consists for all persons only of items that nobody responds to in another session, the model is not identified without additional assumptions (such as that these items are on average equally difficult as the other items). Furthermore, the model accommodates an unbalanced design, that is, all persons need not have the same number and/or length of study sessions—a typical situation in e-learning environments where there is user freedom.

In case the second term $\sum_{k=1}^{K}\left(\alpha_{1k}+\omega_{1pk}\right) stime_{pi} D_{pik}$ of Equation 3 is dropped, and an assumption is made that item scores of an individual in the same study session are item responses of one measurement occasion, the resulting model can be looked at as being equivalent to the multidimensional Rasch growth model of Andersen (1985) except that the item difficulty parameters are defined as random effects in this model. In such a case, growth (or change in ability) is assumed to have occurred between but not within study sessions—which is not the focus of the present article, and as such, will not be pursued any further.

The linear growth assumption of Equation 3 may not be realistic in some situations, for example over longer time periods. If enough observations are made, other growth alternatives can nonetheless be considered. For example, in case quadratic growth is considered a realistic assumption, session specific quadratic terms, $\sum_{k=1}^{K}\left(\alpha_{2k}+\omega_{2pk}\right) stime_{pi}^{2} D_{pik}$, can be included in the model. In this case, the additional terms are α_{2k}—the overall slope for quadratic term in session k and, ω_{2pk}—the deviation of person p within session k from α_{2k}. In this way, the linear term α_{1k} measures the average growth rate in session k, while the quadratic term α_{2k} characterizes the acceleration over time within session k, although the two parameters are not independent of each other. Of course, other non-linear functions like exponential or piece-wise growth can be considered and investigated. The consequence of these extensions is, however, that resulting models may become more complex and difficult to estimate and interpret. Equation 3 (or its extensions to nonlinear forms) can be considered as explanatory generalized (non)linear mixed models as discussed by De Boeck and Wilson (2004), with the (non)linear time trends explaining rates of change in ability or probabilities of success within specific study sessions, and with the random terms modelling the unexplained variability in the success probabilities. Further extensions are possible. In general, the models simply map a substantive theory onto the space of the observed data. The statistical model will not approximate the data well if the assumed theory or distribution is not a good fit to the sample space of the data. As such, alternative model formulations can be fitted to the dataset at hand and model fits compared in order to obtain the most parsimonious model.

Example 1: The Statistics-Online Environment

A tracking and logging sample dataset obtained from a Statistics-Online learning environment was used to illustrate the presented model (Equation 3). Designed and implemented by the authors, Statistics-Online is an item-based e-learning environment for students of the Educational Sciences and the Speech Therapy and Audiology Sciences bachelor programs of University of Leuven in Belgium. Rolled out in November 2011, the environment was designed to supplement classroom based training and to encourage students to practice and improve their ability in statistics. At present, the environment consists of 145 multiple choice items obtained from previous years'

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

TABLE 2
Frequencies of Person Characteristics for the Statistics Online Data

Property	n	%	
Course offered			
Educational Sciences	50	78.1	
Speech and Audiology	14	21.9	
Amount of time spent within study sessions (hours)			
	Minimum	Median	Maximum
Session 1	0.02	1.07	7.88
Session 2	0.05	1.43	10.34
Session 3	0.06	0.95	8.77
Time between study sessions (days)	0.39	2.01	68.40

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

examination questions of one module—regression analysis. For the present article, all analyses in this example are only based on data from the regression module. It is, however, envisaged that other modules like analysis of variance will be added in the e-learning environment in due course. The environment's usage is voluntary much as students are encouraged and motivated by the authors to independently engage the exercises online. Persons are allowed to log in at different times of their choice outside the traditional classroom based lecture times. Upon logging into the environment, persons are presented with all exercise items in a random order, and the possible answers to an item are randomly shuffled for each person. If persons correctly or incorrectly answer an exercise item, they get a detailed feedback on why the selected answer is correct or wrong in order to enhance learning. Whereas the e-learning environment is open to students throughout the entire semester, each item can be engaged by an individual only once.

The considered dataset was collected after two semesters of the rollout phase (i.e., only for the 2011/2012 academic year). Only those persons who have scores on at least five items, engaged within at least two study sessions (64 persons in total) were considered for analysis. Item scores are binary coded (correct/wrong). The total number of exercises engaged in each study session varies considerably across persons. The median number of items per person per session was 13, with a mean of 20 and a maximum of 124. Of the 64 persons, only 1 had up to seven sessions, 4 had up to five sessions, 10 had up to four sessions and 23 had up to three sessions. Overall, individual persons engaged between 10 and 145 items and the distribution of the number of engaged items is skewed to the right with a median and mean of 41 and 51 items respectively. The minimum number of persons per item was 5, with a median of 29 and a maximum of 42 persons. Continuous variables showing time points in hours within specific study sessions on which individual items were engaged by persons were computed as discussed in the "Data Structure" section. Descriptive statistics are given in Table 2. The lowest amount of time spent in a study session by a person was .02 hr (in the first session) and the maximum was 10.34 hr (in the second session).

For a fair comparison of estimates across study sessions for this part of the analysis, only the first 3 study sessions were considered for the present analysis. This is because the persons in the latter sessions are too few (10 persons with four sessions, 4 with five sessions, and only 1 with more than five sessions) such that the estimates of the variance in abilities of persons within these sessions may be very unreliable.

We fitted two longitudinal IRT models to the described Statistics-Online sample dataset to explore changes in persons' ability within specific study sessions: (a) a quadratic model with fixed session-specific average intercepts, linear and quadratic slopes in addition to variances of their respective random person residuals, and (b) a linear growth mode as described in Equation 3. Parameter estimation was done using the *lmer()* function in the *lme4* package (Bates & Maechler, 2011) of R statistical software. The Akaike's information criterion (AIC) and the Schwartz's information criterion (BIC), which are readily obtained from the *lmer()* output, were used to select the best fitting model (see the "Discussion" section for supporting reasons). Fitting these models to the data using the *lmer()* function gives us insight in population-level model parameters or in model parameters of subpopulations of persons. In case individual-level inferences are of interest, the random effects of individual persons (or items) can be estimated using empirical Bayes approaches (cf. Hedeker, Mermelstein, & Flay, 2006). To get a reliable estimate for an individual student, however, a relatively large number of observations for the student is needed, but this issue is not in the scope of this article. A macro used for fitting some of the discussed models is given in Appendix for interested readers.

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

RESULTS

The respective AIC and BIC values are 3709 and 4040 for the quadratic model versus 3666 and 3835 for the linear model, and hence we chose for the linear growth model. However, looking at the complexity of these models and the number of parameters to estimate thereof in relation to the size of this data sample, there is a possibility that these results may not be very reliable, and therefore, we need to exercise caution when making interpretations. As such, presented results from the fitted linear growth model are simply used in the present article for illustration purposes because the number of parameters to estimate from this model are too many to be efficiently estimated from the considered sample size.

The estimated session-specific parameter estimates from the linear growth model are shown in Table 3. It can be noted that the mean ability at the start of each study session depends on the session, and that each session has a different average positive slope estimate. According to this model, the expected probability of an average person giving a correct answer on an average item at the start of the first study session is .49. This is the antilogit of –.06, the intercept for study session 1. However, it can be noted that for the second study session, the expected probability of an average person giving a correct answer on an average item has increased to .59. This suggests that there was an average positive growth in persons' abilities between the start of study session 1 and the start of study session 2, and this change could either be as a result of learning by engaging items within study session 1 but can also be attributed to unobserved persons' activities between study sessions 1 and 2 (see next section). This is confirmed by a chi-square test for the equality of the expected probabilities for study sessions 1 and 2, which was rejected with $\chi^2 = 4.27$, $df = 1$ and $p = .04$. In order to find out whether engaging items within study session 1 led to learning and therefore contributed to this growth in ability, we look at persons' slopes within session 1. It can be noted that on average, the probability of giving a correct answer tends to increase with increase in time within study session 1 because the average slope is positive and statistically significant ($estimate = .12$, $z = 2.21$, $p = .03$). That is, after one hour of making exercises within study session 1, the probability of giving a correct response has increased for .49

TABLE 3
Parameter Estimates of the Multivariate Longitudinal Item Response Theory Models for the Statistics Online Data

Parameter	*Est.*	SE					
Fixed effects							
Intercepts							
Session 1	−.06	.16					
Session 2	.36	.17					
Session 3	−.01	.22					
Slopes							
Session 1	.12	.05					
Session 2	.01	.04					
Session 3	.12	.07					
Random effects	Var.				Correlation		
Persons			S1i	S2i	S3i	S1r	S2r
Intercepts' residuals							
Session 1 (S1i)	.42						
Session 2 (S2i)	.63		1				
Session 3 (S3i)	.50		.83	.83			
Slopes' residuals							
Session 1 (S1r)	.005		.93	.93	.57		
Session 2 (S2r)	.0003		.99	.99	.72	.98	
Session 3 (S3r)	.010		.40	.40	−.18	.71	.55
Items	1.14						
Fit statistics							
AIC	3666						
BIC	3835						

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

to $\text{logit}^{-1}(-.06 + .12^{*}1) = .51$. Similarly, changes in average probabilities of success between and within the other study sessions can be obtained. For the analyzed data sample, average slopes for these subsequent study sessions were not statistically significant at the 5% alpha level. This implies that there is convincing evidence that, on average, persons were learning within the first study session, but there is no convincing evidence for learning during the subsequent ones.

Furthermore, it can be noted from Table 3 that the probability of a correct response seems to vary over persons and over study sessions. For the first study session, the proportion of variance attributed to differences between persons is 8.6%. This is computed as $0.42/(0.42 + 1.14 + \pi^2/3)$ for a two-way random effects model (McGraw & Wong, 1996) with $\pi^2/3 \approx 3.29$ equal to the variance of level 1 residuals for a standard logistic distribution (Snijders & Bosker, 1999). Similarly, the proportion of variance attributed to differences in persons for study sessions 2 and 3 are 12.4% and 10.1% respectively. It is noted from Table 3 that although persons tend to have differences in growth trends (rates of change in ability) of success probabilities within individual study sessions, these differences between person slopes seem to be minimal. Note also from Table 3 the very high and positive correlation values between study sessions' intercepts and slopes. Although, as mentioned earlier, we much be very careful in interpreting the parameter estimates of this complex model given the small number of observations, this is an indication that

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

those persons who performed highly in a study session, tended to perform well in subsequent study sessions, and that over the sessions, the same persons learn relatively fast. The proportion of variance attributed to differences in items for each study session can also be computed and these as expected are relatively similar, that is 23.6%, 22.6% and 23.2% for study sessions 1, 2 and 3 respectively.

Measuring Growth Within and Between Study Sessions

During the time between study sessions, there may be some constructive learning, for example by studying printed materials or following a course outside the e-learning environment but there can also be negative learning like forgetting what one has already learned. Although such information is not tracked and logged into person specific trails, the amount of time spent outside the environment between two consecutive study sessions can be obtained (refer to Table 1). It may therefore be desirable that Equation 3 be extended to obtain rates of growth that differ per person, per study session and per between-study session period. In this way, a person's ability can be assumed to evolve over time—for within-study sessions possibly because of the learning that takes place, and between-study sessions as a result of the person's unobserved activities. Therefore, we could extend Equation 3, by including not only slopes for the study sessions, but also for the intervals between study sessions. A caveat of this model (as well of Equation 3) is that in some e-learning environments, the number of study sessions per person can be rather large either because of study design or user freedom. In the study of Klinkenberg and colleagues (2011), for example, it is reported that each pupil was required to complete 15 items in a single session, yet the item bank contained 738, 723, 659, and 664 items in the respective domains of addition, subtraction, multiplication, and division. In such a case, the number of sessions and therefore the number of parameters to estimate from Equation 3 can be very large, thereby limiting easiness in interpretation of results. Moreover, study sessions might be too short to obtain reliable parameter estimates. Furthermore, because of learner freedom in some e-learning environments, the number of items per session over persons is likely to vary greatly, as well as the space between study sessions. This makes that the comparability of study sessions over persons questionable. For example, if a person responds to 66, 1 and 112 items in three consecutive sessions, the responses for the 112 items will contribute to the estimate of the overall initial ability and slope for the third session, although it can be expected that there was not much learning in the very short second session for this individual and therefore it would make more sense to regard the session with the 112 items as the second session.

To overcome some of these challenges, a few constraints can be introduced to Equation 3 and its extension with between study session slopes. We can, for example, assume that there are differences between persons in the initial ability and in the learning trends, but that for a given person, the rate of change in ability is the same within all study sessions, and the same within all between-study sessions. In this way, the study session time indicators in Equation 3 are constrained such that there is only one growth parameter estimate per person, one for all within-study sessions. The same can be done for all between-study spaces. Both slopes are modeled by computing a within-session time variable (*wtime*) and a between-session time variable (*btime*) from the real-time points within and between study sessions as given in the last two columns of Table 1. The variable *wtime* keeps increasing over all study sessions of an individual but pauses between sessions. Similarly, the variable *btime* increases between sessions until the point when

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

an individual commences using the environment again. Thus, the probability (π_{pi}) of a correct response by person p to item i can be given as follows:

$$\text{logit}\left(\pi_{pi}\right) = (\alpha_0 + \omega_{0p}) + (\alpha_1 + \omega_{1p})wtime_{pi} + (\alpha_2 + \omega_{2p})btime_{pi} + \upsilon_i, \quad (4)$$

where α_0 is the overall initial ability, ω_{0p} is the deviation of person p from α_0, $wtime_{pi}$ and $btime_{pi}$ is the amount of time that passed for person p while respectively using and not using the learning environment, up to the moment person p responds to item i, α_1 and α_2 are overall population linear time trends within and between sessions respectively, ω_{1p} and ω_{2p} are deviations of person p from α_1 and α_2 respectively. As before, the person random effects can be assumed to have a multivariate normal distribution. Furthermore, it can also happen that growth does not strictly occur linearly within and between study sessions as presented in Equation 4. If, for example, we assume that growth occurs quadratically within all study sessions, Equation 4 can be extended by including a quadratic term, $(\alpha_3 + \omega_{3p})wtime_{pi}^2$, where α_3 is the average slope for quadratic time trends within study sessions, and ω_{3p} is deviation of persn p from α_3. Similarly, a quadratic term for the between study sessions variable can be included.

Explanatory Extensions

As in other educational settings, data from e-learning environments can contain additional design structures. For example, there can be observed person properties that may influence changes in the persons' ability. As such, person differences in growth might partly be explained by person properties, such as gender, course offered, or school. In this regard, the presented generalized linear mixed models that may be extended to an explanatory IRT framework (De Boeck & Wilson, 2004) by including other predictors. The explanatory IRT modeling framework is very flexible as it can be extended to different IRT models explaining latent properties of items, of persons and of their groups.

Suppose that persons are grouped by a categorical property with levels $g = 1, \ldots, G$. The linear model in Equation 4, for example, can be extended to measure group differences in growth (group specific changes in ability), such that, the probability (π_{pgi}) of person p belonging to group g to correctly answer item i can be obtained as follows:

$$\text{logit}\left(\pi_{pgi}\right) = (\alpha_0 + \omega_{0p} + \tau_{0g}) + (\alpha_1 + \omega_{1p} + \tau_{1g})wtime_{pi} + (\alpha_2 + \omega_{2p} + \tau_{2g})btime_{pi} + \upsilon_i, \quad (5)$$

with τ_{0g}, τ_{1g}, and τ_{2g} random group effects, respectively referring to the deviations of the intercept, growth during the use of the learning environment and the growth in between using the learning environment, from the overall intercept and growth parameters. These random effects are assumed to follow a multivariate normal distribution with zero means. If the number of groups is relatively small and/or if groups cannot be regarded as a randomly chosen sample of groups, it might be more adequate to use group properties as predictors in the model. For example, if we have two groups, we could use the following model:

$$\begin{aligned}\text{logit}\left(\pi_{pi}\right) = {} & (\alpha_{01}Group1_p + \alpha_{02}Group2_p + \omega_{0p}) \\ & +(\alpha_{11}Group1_p + \alpha_{12}Group2_p + \omega_{1p})wtime_{pi} \\ & +(\alpha_{21}Group1_p + \alpha_{22}Group2_p + \omega_{2p})btime_{pi} + \upsilon_i, \end{aligned} \quad (6)$$

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

with *Group*1 and *Group*2 dummy variables which equal 1 if person p belongs to respectively the first and the second group, zero otherwise. The parameters α_{01}, α_{11}, and α_{21} now can be interpreted as the expected initial level and growth in the first group, α_{02}, α_{12} and α_{22} as the expected initial level and growth in the second group.

Equations 5 and 6 can be looked at as person explanatory IRT models (De Boeck & Wilson, 2004) that explain differences in person ability without assuming that the person predictors explain all the variance. Moreover, other available person properties that may not be categorical, for example age, can be included in the models as well. By including random person effects and person predictors as fixed effects, individual differences can be assessed and relations with measures of theoretically related constructs can be postulated, which is a way of assessing external validity (Spence, Owens, & Goodyer, 2012). Note that the item difficulty parameters remain the same for groups in Equations 5 and 6, but it is also possible to include interaction terms of person groups by item difficulty parameters in the model to assess presence of differential item functioning following procedures in the study of Van den Noortgate and De Boeck (2005). This, however, is not considered in the present study. Although not considered further in this study, the multisession longitudinal model in Equation 3 can similarly be extended to this explanatory IRT approach.

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

Examples

Two datasets were analyzed to illustrate the longitudinal IRT models for measuring growth within and between study sessions—that from the Statistics-Online learning environment, and another from the Maths Garden learning environment (Klinkenberg et al., 2011). Although datasets collected from the environments are generally similar in structure, they differ in person and item characteristics as discussed in respective sections. For both datasets, the following longitudinal IRT models were fitted and compared:

1. A general quadratic model with average intercepts, linear and quadratic slopes both within and between sessions, and variances of respective random effects.
2. A linear model with a fixed average intercepts and only linear slopes within and between sessions, and variances of respective random effects (this is Equation 4).

As before, parameter estimation was done using the *lmer()* function in R statistical software. Obtained AIC and BIC values were used to select the best fitting model to the present dataset. The selected model was then extended to the explanatory framework as discussed in Equation 6 with a person property of interest. A macro used in model fitting is provided in the Appendix for interested readers.

Example 1: The Statistics-Online Environment

The dataset from the Statistics-Online learning environment is already described under Equation 3. Recorded in this dataset was a person property *Course* indicating the two bachelor programs that persons were following, and this was used to illustrate the explanatory longitudinal IRT model by means of two dummy variables, called *Speech* and *Education*. Table 2 shows that the largest proportion of persons (78.1%) belonged to the Educational Sciences group. The continuous

variable *wtime*, showing measurement time points in hours across all study sessions; and *btime*, showing the spacing time in days between study sessions were computed from the session specific time stamps. The median spacing time between two consecutive study sessions was two days, the maximum 68.4 days (Table 2).

RESULTS

For this dataset, the obtained AIC and BIC pairs for the respective fitted models are 3925 and 4053 for the quadratic model versus 3872 and 3933 for the linear model. From these values, the linear slopes model seems to yield the best balance between parsimony and fit. As noted before, however, the stability of these results can be questioned because of the small sample size and we emphasize here that these results are given for illustration purposes.

Obtained results from the linear model are shown in Table 4. From Model 4 in this table, we learn that the expected probability of an average person giving a correct answer on an average

TABLE 4
Estimates of the Parameters of the Item Response Theory Models for the Statistics Online Environment

	Model 4		*Model 6*				
Parameter	*Est.*	SE	*Est.*	SE			
Fixed effects							
Intercept	.12	.15					
Speech			.51	.21			
Education			−.03	.17			
Within-sessions slopes	.04	.02					
Speech			.05	.06			
Education			.05	.02			
Between-sessions slopes	.00	.00					
Speech			.02	.01			
Education			.00	.00			
Random effects	Var.		Var.		Correlation		
Persons				S	E	WS	WE
Intercept	.53						
Course							
Speech (S)			.22				
Education (E)			.56				
Within sessions	.0004						
Speech (WS)			.0008	1			
Education (WE)			.0008		.95		
Between sessions	.0001						
Speech (BS)			.0001	−1		−1	
Education (BE)			.0002		−1		−.95
Item	1.14		1.14				
Fit statistics							
AIC	3872		3894				
BIC	3933		4064				

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

item before any learning takes place (i.e., at time zero) is .53 which is the antilogit of .12. It can be noted, however, that on average the probability of giving a correct answer tends to increase within study sessions, reflected by the positive and significant parameter estimate associated with the *wtime* covariate ($estimate = .04, z = 2.14, p = .03$). After 1 hr of making exercises, the probability of giving a correct response increases to $\text{logit}^{-1}(.11 + .04) = .54$. In this data sample, the obtained parameter estimate for the slopes between study sessions is almost zero and not significant ($estimate = .00, z = 0.70, p = .48$). As in the analyses using Equation 3, the proportion of variance in probabilities of success attributed to differences in persons and items can be computed and these are 10.8% and 23.0% respectively. It can be further observed that there is some variability in differences between persons in the growth trends within study sessions ($var = .0004$) and between study sessions ($var = .0001$).

There is a very high positive correlation between the intercept deviations (ω_{0p}) and the within sessions slope deviations (ω_{1p}) for $\rho_{\omega_o\omega_1} = 1$ (not shown in Table 4). On the contrary the correlation between ω_{0p} and the random deviations of the slope between sessions (ω_{2p}), and that between ω_{1p} and ω_{2p} were high and negative, $\rho_{\omega_o\omega_2} = \rho_{\omega_1\omega_2} = -1$. This may be an indication that those persons who started with large probabilities of success, tended to learn faster during sessions and have a smaller slope between sessions. The very high correlation coefficients may also be an indication that the results are not too stable, perhaps because of the small sample size or are influenced by some extreme cases. Note that in computing the correlation coefficients, it is possible for the product of the standard deviations to be larger than the covariance of interest and in such a case, obtained values are truncated to 1.

Model 4 gives general observations for all persons participating in the learning environment. However, by including the person property *course* as a predictor, it can be noted from Model 6 (Table 4) that there are group differences in growth. Whereas persons following the bachelor of speech and audiology sciences have an average probability of .62 to succeed on average items before any use of the learning environment, the one for persons following the bachelor of education sciences is .49. A test for equality of both groups at time zero resulted in $\chi^2 = 4.75$ with $df = 1$ and $p = .03$ meaning that there is a statistically significant difference between persons following the two courses. We can, further, observe that growth for the latter group seems to be occurring within study sessions ($estimate = .05, z = 2.37, p = .02$) but not between sessions ($estimate = .00, z = 0.05, p = .96$). On the contrary, the former group of persons seems to have significant growth in ability between study sessions ($estimate = .02, z = 2.64, p = .01$). Within study sessions, the slope estimates for both groups are very similar, but not significant for the first group ($estimate = .05, z = 0.87, p = .38$). This implies that growth in persons' ability may not only be a result of engaging items within the e-learning environment but also between the e-learning study sessions, for example by participating in lectures, discussing what they had done during sessions with peers or studying other printed material. Further from Model 6 of Table 4, it is observed that there was more variability in the initial abilities of persons following the bachelor of education sciences ($var = 0.56$) compared with those following the bachelor of speech and audiology sciences ($var = 0.22$)—although this could also be due to the small sample size of the later group. Similar group-specific observations can be made for variances in growth within and between study sessions. Much as Model 6 has larger AIC and BIC values as shown in Table 4 compared with those of Model 4 for this data sample, and therefore is not necessarily the most parsimonious model when the two are compared, it can be noted that by fitting an explanatory longitudinal IRT model, one is able to make group-specific

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

inferences and consequently make more informed instructional decisions with the purpose of either improving the learning environment, or enlisting group-specific solutions to improve student performance.

Example 2: The Maths Garden Environment

A sample dataset from the Maths Garden environment (Klinkenberg et al., 2011) was also analyzed for comparability purposes. This dataset has a similar structure like that from the Statistics-Online environment, but differs in the implementation methodology in to having a larger number of participating persons and engaged items. The Maths Garden is a web-based computer adaptive practice and monitoring system used in The Netherlands to provide more time to practice and maintain basic arithmetic skills in order to improve the ability of individual students. Persons can regularly engage specific exercise items that are adjusted to their individual ability levels using an IRT model and also includes immediate feedback to persons on the exercises they engage thereby ensuring that learning takes place. According to Klinkenberg and colleagues (2011), the Maths Garden was especially developed to improve mathematics education in primary schools by combining arithmetic practice and measurement in a playful manner using computerized educational games. In their article, for example, the analyzed dataset for the 2008/2009 academic year consisted of pupils from 35 schools from across The Netherlands within the age range of 4–12 years enrolled in Kindergarten and within 1–6 grades of primary school. In the present article, however, the analyzed dataset for the 2010/2011 academic year consists of item responses of the participants from 177 schools from across The Netherlands, aged 4–18 years with the additional older participants enrolled in the first 4 years of secondary school education. For a detailed design and implementation methods of the Maths Garden e-learning environment, the interested reader is referred to Klinkenberg and colleagues (2011).

Four domains of addition, subtraction, multiplication and division are administered within the Maths Garden environment, but the analysis of this article focuses only on the former domain—addition. The analyzed dataset consists of response scores to 610 items by 3803 persons. The analysis in this article focuses on individual persons who engaged at least 15 exercise items. The highest number of engaged items by an individual was 227 with a median and mean of 49 and 53 exercise items, respectively. As in the case of the Statistics-Online environment, the recorded time stamps were used to compute a person's study sessions and the highest number of sessions by an individual was 33. Furthermore, the variables *wtime* in hours and *btime* in days for each person were also computed from the time stamps. For this dataset, the minimum study session length by a person was .01 hr with a median of .16 and a maximum of .68 hr. Furthermore, the minimum spacing time was one day with a median of five and a maximum of 53 days respectively. Although other properties exist in this dataset, only one person property gender (male = 53.6%, female = 46.4%) was used to illustrate the explanatory longitudinal IRT model. The response variable is the first attempt score on an item by a person with 1 = pass and 0 = fail.

RESULTS

For this example, the quadratic model reached iteration limit before convergence could be achieved. This is perhaps because of the small number of engaged items for persons (mean

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

TABLE 5
Estimates of the Parameters of the Item Response Theory Models for the Maths Garden Environment

	Model 4		*Model 6*				
Parameter	*Est.*	SE	*Est.*	SE			
Fixed effects							
Intercept	1.01	0.01					
Female			0.93	0.02			
Male			1.08	0.02			
Within-sessions slopes	0.81	0.11					
Female			1.17	0.18			
Male			0.52	0.14			
Between-sessions slopes	−0.01	0.00					
Female			−0.01	0.00			
Male			−0.01	0.00			
Random effects	Var.		Var.		Correlation		
Persons				F	M	WF	WM
Intercept	0.21						
Gender							
Female (F)			0.23				
Male (M)			0.19				
Within sessions	6.71						
Female (WF)			8.80	−0.63			
Male (WM)			5.42		−0.56		
Between sessions	0.0002						
Female (BF)			0.0003	−0.23		−0.41	
Male (BM)			0.0002		−0.46		−0.13
Items	0.014		0.014				
Fit statistics							
AIC	234556		234541				
BIC	234659		234828				

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

of 53 out of 610 items) such that for this complex model, the number of parameters to estimate require more information than the data can support. Therefore, only the model with linear slopes within and between study sessions was considered.

Obtained results by fitting this model to the Maths Garden dataset are shown in Table 5. For this dataset, all the average growth slopes within and between study sessions are statistically significant at a 5% alpha level. To understand the growth of persons' abilities over time, we look at the signs of the overall slope estimates of this model. Whereas the average slope between sessions is negative ($estimate = -0.01, z = -8.58, p < .001$), the one within sessions is positive ($estimate = 0.81, z = 7.20, p < .001$). This is an indication that the probabilities of success for persons, which can be computed as was illustrated in Example 1, will on average tend to increase within study sessions, and decrease between study sessions.

In this dataset, the proportion of variance attributed to differences between persons at the start of using the learning environment is larger (6.1%) when compared with that due to differences between items (0.4%). It is also observed that there are differences between persons in the evolutions of success probabilities within and between study sessions. There are generally negative

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

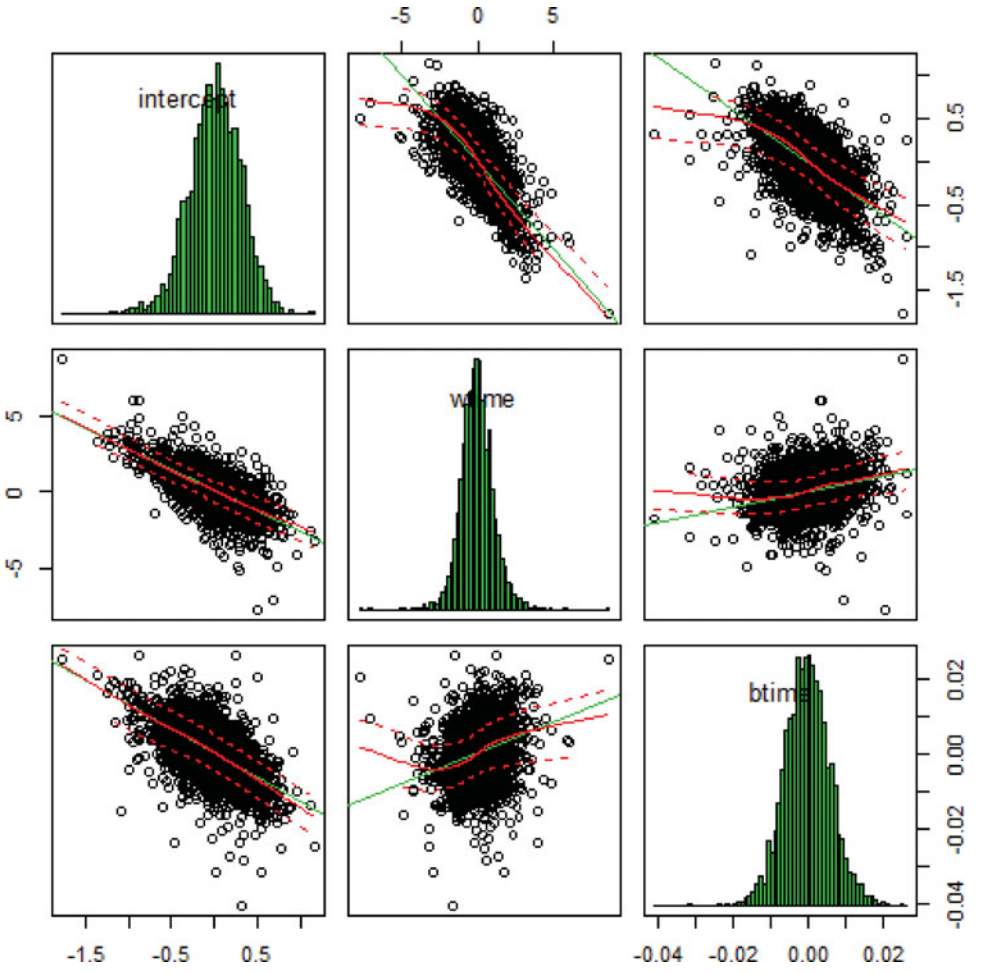

FIGURE 1 Scatter plot matrix for the random effects estimates for the Maths Garden Environment.

correlation coefficients between the intercept deviations (ω_{0p}), the random deviations from the slope within sessions (ω_{1p}) and the random deviations of the slope between sessions (ω_{2p}) respectively given as $\rho_{\omega_o\omega_1} = -.60$, $\rho_{\omega_o\omega_2} = -.38$, and $\rho_{\omega_1\omega_2} = -.22$ (not shown in the table). This implies that persons who started with large probabilities of success, tended to have a smaller slope during sessions and a smaller slope between sessions.

The scatter plot matrix in Figure 1, obtained from the random effects' estimates of Model 4, gives us an idea of the plausibility of the assumption of multivariate normality between the intercept deviations (ω_{0p}), the within sessions slope deviations (ω_{1p}) and the random deviations of the slope between sessions (ω_{2p}). In Figure 1, it is noted that the univariate distributions for each variable (on the diagonal) are approximately normal. Similar observations were made for the same model fitted to the sample dataset in Example 1, although the bivariate normality

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

for the off-diagonal elements was not to be well met, perhaps due to the small samples in that example. However, just like in this example, the Q-Q plots (not shown) for each of the person variables (ω_{0p}, ω_{1p}, ω_{2p}) and that of the random item difficulties (υ_i) were also used to check for univariate normality assumptions, and in each example, the linearity of the points on the diagonal axis suggested that all these variables were generally more or less normally distributed.

Including gender in the Model 6 allows one to obtain and make inferences about the differential growth trends for men and women both within and between study sessions. It can, for example, be observed from Table 5 that compared with men, women have on average higher growth over time (i.e., the slope) both within and between study sessions; and this difference is statistically significant with $\chi^2 = 2949.3$ with $df = 18$ and $p < .001$. As before, proportions of variance in differences of success probabilities attributed to differences in persons per group can be computed. For this dataset, the AIC and BIC values shown in Table 5 do not agree on the preferred model (Model 4 with no person covariate versus Model 6 with a person covariate). For AIC, the most complex (quadratic) model is "best" while for BIC the one with fewer degrees of freedom (the linear model) is "best." As before, the advantage of Model 6 is that it provides group-specific growth estimates which is not the case with Model 4.

DISCUSSION

This article presented longitudinal IRT models for measuring growth over time in item-based tracking and logging data from e-learning environments. The presented models are, in fact, extensions of the longitudinal IRT models (cf. Andersen, 1985; Embretson, 1991) which reveal individual differences in ability change when item sets are presented to persons over two or more measurements occasions—equivalent to the study session in this article. The exceptions included in this study are that item difficulty parameters are defined as random effects, and, the sets of items within specific study sessions need not be the same for all persons. The first extension (Equation 3) measures changes in persons' ability (growth) within specific study sessions, that is, a person's ability is modeled as linearly changing over time with session-specific slopes. The second extension encompasses changes in persons' ability both within and between study sessions.

The discussed models were described with hindsight that e-learning environments are likely to continue to proliferate in educational settings and as such, researchers and educationists would be interested in measurement of growth in ability and how it changes across time for persons using such e-learning environments. The described longitudinal IRT models are basically a class of psychometric Rasch models which can be easily extended to model various aspects of growth, such as global growth (for example by looking at the initial and final estimates), session-specific growth for a population, as well as different patterns of growth for different subpopulations (cf. Meiser et al., 1999). In contrast, measuring growth within and between study sessions, can give session specific estimates in addition to providing ability growth estimates at any time during the learning process. In this regard, illustrations using two empirical examples from e-learning environments were given explaining how one can measure population growth over time or group specific growth by including relevant person predictors in the model. For example, a person property in each of the models was included using the explanatory IRT modeling framework

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

based on the logistic item response model formulation followed by De Boeck and Wilson (2004). The explanatory approach in longitudinal IRT models is useful when one is interested in a multiple person group comparison on change and in the change score interpretation linked to item attributes (Cho, Athay, & Preacher, 2012).

For the considered examples, the assumption of linearity was fairly reasonable. As such, illustration results in all cases indicated that the presented linear models can be used to track changes in a person's ability over time either within or between study sessions also for different groups of persons. This is important because one could use the presented models to evaluate the effect of a learning environment on persons' abilities, and to try to explore what factors (e.g., item or person characteristics) moderate this effect or can contribute to this effect. In the Statistics-Online environment, for example, it was observed that one group of persons following the bachelor of Speech and Audiology Sciences seemed to have significant changes in their probabilities of success between but not within study sessions, while those following the bachelor of Educational Sciences seemed to have significant changes in their probabilities of success within but not between study sessions. In such a case, an instructor can utilize the results to put in place group-specific procedures to improve students' ability and consequently, their future performance. Although the discussed results focus on the interpretation of the population-level model parameters, individual level inferences (or person specific growth) can be estimated by extracting random effects' estimates by empirical Bayes approaches and adding these to the fixed effects' estimates of a given model of interest.

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

Parameter Estimation

The discussed models were presented within a framework of generalized linear mixed models and can all be fitted using open source software R *lmer* function. The R lmer function has been used for various item response models with crossed random effects (cf. De Boeck et al., 2011). Parameter estimation for IRT models and generalized linear mixed models typically involves maximum likelihood or restricted maximum likelihood (REML; Hedeker et al., 2006). Unfortunately, direct (RE)ML estimation of models for categorical data is unfeasible because the marginal likelihood does not have a closed form, thereby requiring estimation with numerical or Monte Carlo integration (Cho et al., 2012). Instead, parameter estimates under the *lmer()* function can be obtained under (RE)ML with the Laplace approximation of the likelihood. The Laplace approximation to the likelihood is optimized to obtain approximate values of the (RE)ML estimates for the parameters of interest and for the corresponding conditional modes of the random effects (Doran, Bates, Bliese, & Dowling, 2007). In random effects models, the (RE)ML estimates of variance components are generally preferred because they are in general less biased than maximum likelihood estimates (Bates, 2008; Sarkar, 2008). However, Laplace approximation can perform poorly for dichotomous responses with small number of items, resulting in overestimated covariances of the covariance matrix of the random effects (Cho et al., 2012). This may explain the perfect correlation coefficients for the Statistics-Online results in . In terms of runtime, the *lmer()* function is known to be generally fast when compared with other procedures that are based on the similar statistical background like SAS PROC NLMIXED (cf. De Boeck et al., 2011). Run time, however, is influenced by factors like complexity of model in question, size of the dataset (that is, number of items and number of persons), and capacity of the computer used. For example, fitting the session-specific linear model in Equation 3 to the

statistics online dataset takes approximately 83 seconds to achieve convergence on a computer with a CPU of Intel Core i5 with 2.67 GHz and RAM 4GB. On the contrary, the simpler linear model in Equation 4 takes about 22 seconds on the same computer. These run-times do increase when fitting explanatory extensions of the respective models and/or their quadratic versions, and when fitted to the Maths Garden dataset that is relatively larger.

Furthermore, although we adopt an empirical Bayes approach in the present article, the described models can also be fitted using Bayesian inference via Markov chain Monte Carlo estimation algorithm in WinBUGS statistical software (Lunn, Thomas, Best, & Spiegelhalter, 2000). In this respect, a WinBUGS macro to fit the model in Equations 3 and 4 to the sample data from Example 1 is provided in the Appendix for interested readers and obtained results were not very different from those obtained under the empirical Bayes approach presented in this article. Macros for other discussed equations can easily be derived from those provided.

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

Model Fitting

Some of the presented models are nested in each other. Nesting here means that constraining one or more parameters in a more general model, most often to zero, results in the more restricted model (De Boeck et al., 2011). In the present study, the longitudinal IRT models (Equations 3 and 4) are nested in their corresponding explanatory counterparts. For example, Equation 4 can be obtained from Equation 5 by constraining parameters of the person groups to zero. A classical statistical test for the comparison of nested models with different mean structures is the likelihood ratio test (Verbeke & Molenberghs, 2000). However, because the models are fitted using the (RE)ML rather than the maximum likelihood estimation procedure, the likelihood ratio test is no longer suitable for testing differences in the fixed part (Pinheiro & Bates, 2000; Verbeke & Molenberghs, 2000). The Akaike's information criterion, AIC, and the Schwartz' s information criterion, BIC which are readily obtained from the *lmer()* output however still can be used to select the best fitting model. In this study, AIC and BIC were used to compare the quadratic and linear models in order to select the model that described the available data better. Moreover, these criteria were also used in Tables 4 and 5 to compare a longitudinal IRT model in Equation 4 to its explanatory extension (Equation 6) by examining the significance of the additional model parameters.

Limitations

The longitudinal IRT model presented in Equation 3 which measures growth specific to study sessions is of much interest to this study but particularly preferred when the number of study sessions is relatively small. In case the number of study sessions is moderately large, the model parameters to estimate can be many, thereby limiting easiness in interpretation. There is also a danger of over-fitting the data in case some study sessions have few observations and, as explained above, when the study sessions may not be comparable because of personalization of some e-learning environments. This model is suitable for datasets that are fairly balanced with study sessions containing a representative (and preferably comparable) number of persons. In case the person samples in subsequent study sessions are not comparable to those of previous sessions,

variance estimates in person abilities across sessions may be incomparable and can perhaps lead to wrong conclusions. Some of these weaknesses are partly covered by the longitudinal IRT model in Equation 4 and its explanatory extension (Equation 5). In this model, each person is assumed to have only one linear slope within all study sessions and one linear slope between all study sessions. This, however, can be considered a strong and restrictive assumption that may not apply in some practical settings. For example, given the results of Equation 3 in Table 3, it can be noted that between the sessions, the ability seems to go up and down, rather than being constantly increasing/decreasing over time. It should however be noted in this case that the presented models are relatively complex for applying on such a small dataset because we have a large covariance matrix, given the number of students. Nonetheless, it would be interesting for future research to employ and test further the presented models and extensions (for example models with non-linear slopes) with larger data samples having no or less missing scores compared with those considered in the present study or to simulated data.

The models further assume simple distributional assumptions such as multivariate normality of the person random effects. There is need for these assumptions to be formally tested and confirmed in future research, possibly with additional data samples or with simulated data. For example, though not applied in the present article, Mardia (1974) proposed tests that are based on definitions of multivariate skewness and kurtosis can be used as goodness-of-fit tests for multivariate normality. Another limitation of the presented approach is the measurement of the within-session time variable (*wtime*) and a between-session time variable (*btime*) from the real-time points within and between study sessions. In the present article, it is assumed that during an individual's study session, the individual is continuously making use of the learning environment. This, however, may be misleading because a person may take a break, for example, to make a telephone call or to have a snack without necessarily logging out of the learning environment. Yet this is not captured by the designs described in the present article. Other definitions of these variables could be explored in future research and/or with other learning environments. Furthermore, caution is especially needed when extreme parameter values are encountered, such as the correlations equal or close to 1 as in the first example. In this respect, future simulation studies could shed light on the number of items and/or persons needed in order to obtain stable estimates. That said, we can expect that in the future much more large scale electronic learning environments will be available, such as Maths Garden.

CONCLUSION

Despite the discussed shortcomings, the presented models achieved the overall objectives of the present study. That is, the models present us possibilities to understand whether persons' abilities are changing within and/or between study sessions and how these changes can contribute to persons' probabilities of success over time in e-learning environments. Several extensions of the presented models can be made. For example, this study did not include any item groups as predictors but these can be easily be handled as in the case of the *item explanatory* or the *doubly explanatory* IRT models (De Boeck & Wilson, 2004). Furthermore, we relied only on observed person properties but there could be unobserved person properties contributing to the changes in persons' ability and these can be taken into account by extending the presented models to, for example, the framework of mixture Rasch models (Rost, 1990). Furthermore, one can include

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

interaction terms of person groups by items as random interaction effects in each of these models in order to assess presence of differential item functioning (cf. Van den Noortgate & De Boeck, 2005). Alternatively, inclusion of interaction terms between person and item groups, or between study sessions and item groups in the model, may provide possibilities of estimating persons' group-specific differences across the different levels of item groups, which can be a possible measure for presence of differential item-group functioning for the different levels of person groups or study sessions. In addition, there is a possibility of guessing by persons within the discussed e-learning environments especially as a result of user freedom and personalization of the study sessions, and all items may not be necessarily equally discriminating. Although Cho and colleagues (2012) noted that the feasibility of estimating item discrimination parameters may be limited in case of small sample sizes, it can nonetheless be interesting to extend the presented models to include guessing and discrimination parameters.

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

ACKNOWLEDGEMENTS

The authors give kind acknowledgements to Prof. Han L. J. van der Maas, University of Amsterdam, and to Oefenweb.nl for providing the dataset from the Maths Garden learning environment; and to Martin Vanbrabant, Marlies Vervloet, Peter Jouck, and Sander Vanhaute for assisting with the design and implementation of the Statistics-Online learning environment.

AUTHOR NOTES

Damazo T. Kadengye is a post-graduate student in the Faculty of Psychology and Educational Sciences, University of Leuven. **Eva Ceulemans**, is a professor at the Centre for Methodology of Educational Research, University of Leuven. **Wim Van Den Noortgate** is a professor in the Faculty of Psychology and Educational Sciences, University of Leuven.

REFERENCES

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

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis*, *95*, 1–22. doi:10.1016/j.jmva.2004.07.005

Bates, D. M. (2008). *Fitting mixed-effects models using the lme4 package in R*. Retrieved from http://www.stat.wisc.edu/∼bates/PotsdamGLMM/LMMD.pdf

Bates, D. M., & Maechler, M. (2010). *lme4 0.999375-42. Linear mixed-effects models using S4 classes [R package]*. Retrieved from http://cran.r-project.org

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning. A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102. doi:10.1111/j.1467-9280.2008.02209.x

Chen, C.-M., Lee, H.-M., & Chen, Y.-H. (2005). Personalized e-learning system using item response theory. *Computers & Education*, *44*, 237–255. doi:10.1016/j.compedu.2004.01.006

Cho, S.-J., Athay, M., & Preacher, K. J. (2012). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *The British Journal of Mathematical and Statistical Psychology*, *66*, 353–381. doi:10.1111/j.2044-8317.2012.02058.x

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.

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

Desmet, P., Paulussen, H., & Wylin, B. (2006). FRANEL: A public online language learning environment, based on broadcast material. In E. Pearson & P. Bohman (Eds.), *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 2307–2308). Chesapeake, VA: AACE. Retrieved from http://www.editlib.org/p/23329

Doran, H., Bates, D. M., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20*, 1–18.

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

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika*, *60*, 459–487. doi:10.1007/BF02294324

Guskey, T. R. (2003). How classroom assessments improve learning. *Educational Leadership*, *60*, 6–11.

Hedeker, D., Mermelstein, R. J., & Flay, B. R. (2006). Application of item response theory models for intensive longitudinal Data. In T. A. Walls, J. L. Schafer, & T. A. Walls (Eds.), *Models for intensive longitudinal data* (pp. 84–108). Oxford, UK: Oxford University Press.

Johns, J., Mahadevan, S., & Woolf, B. (2006). Estimating student proficiency using an item response theory model. In M. Ikeda, K. Ashley, & T.-W. Chan (Eds.), *Intelligent tutoring systems, Lecture notes in computer science* (Vol. 4053, pp. 473–480). Berlin/Heidelberg, Germany: Springer. doi:10.1007/11774303_47

Klinkenberg, S., Straatemeier, M., & Van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*, 1813–1824. doi:10.1016/j.compedu.2011.02.003

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337. doi:10.1023/A:1008929526011

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *The Indian Journal of Statistics: Sankhya, Ser. B*, *36*, 115–128.

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

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Meiser, T., Stern, E., & Langeheine, R. (1999). Latent change in discrete data: Unidimensional, multidimensional, and mixture distribution Rasch models for the analysis of repeated observations. *Methods of Psychological Research Online*, *3*, 75–93.

Miceli, R., Settanni, M., & Vidotto, G. (2008). Measuring change in training programs: An empirical illustration. *Psychology Science Quarterly*, *50*, 433–447.

Nam, C. S., & Smith-Jackson, T. L. (2007). Web-based learning environment: A theory-based design process for development and evaluation. *Journal of Information Technology Education*, *6*, 23–43.

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

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.

Roberts, J. S., & Ma, Q. (2006). IRT models for the assessment of change across repeated measurements. In R. W. Lissitz (Ed.), *Longitudinal and value added modeling of student performance*. Maple Grove, MN: JAM Press.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282. doi:10.1177/014662169001400305

Safer, N., & Fleischman, S. (2005). How student progress monitoring improves instruction. *Educational Leadership*, *62*, 81–83.

Sarkar, D. (2008). *Fitting mixed-effects models using the lme4 package in R*. Retrieved from http://www.bioconductor.org/help/course-materials/2008/PHSIntro/lme4Intro-handout-6.pdf

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, UK: Sage.

Spence, R., Owens, M., & Goodyer, I. (2012). Item response theory and validity of the NEO-FFI in adolescents. *Personality and Individual Differences*, *53*, 801–807. doi:10.1016/j.paid.2012.06.002

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*, 443–464. doi:10.3102/10769986030004443

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*, 369–386. doi:10.3102/10769986028004369

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.

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

Verguts, T., & De Boeck, P. D. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement*, *24*, 151–162. doi:10.1177/01466210022031589

Verhelst, N. D., & Glas, C. A. W. (1995). Dynamic generalizations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 181–201). New York, NY: Springer.

Vlach, H. A., & Sandhofer, C.M. (2012). Distributing learning over time: The spacing effect in children's acquisition and generalization of science concepts. *Child Development*, *83*, 1137–1144.

Wauters, K., Desmet, P., & Van den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, *26*, 549–562. doi:10.1111/j.1365-2729.2010.00368.x

Welsh, E. T., Wanberg, C. R., Brown, K. G., & Simmering, M. J. (2003). E-learning: emerging uses, empirical results and future directions. *International Journal of Training and Development*, *7*, 245–258. doi:10.1046/j.1360-3736.2003.00184.x

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

APPENDIX

R Macro

```
## import data ###
rm(list = ls())
learn <- read.table("statsonline.txt",sep = ";", header = TRUE)
## prepare data for analysis#
str(learn)
learn$time <- as.character(learn$time)
learn$userid <- as.factor(learn$userid)
learn$itemid <- as.factor(learn$itemid)
learn$course <- as.factor(learn$course)
## Retain users with > = 10 items ##
count <- function(x) {
      length(na.omit(x))}

ci.userid <-
data.frame(aggregate(learn$item,by = list(learn$userid),FUN = "count")) # count of items per user
colnames(ci.userid) <- c("userid","citem")
learn <- merge(learn,ci.userid,by = "userid")
learn <- subset(learn, citem%in% citem[which(citem> = 10)])
learn$citem <- NULL; rm(ci.userid)

#### Compute user-attempt time lengths and create study sessions ####
ct <- learn$time
ctparts = t(as.data.frame(strsplit(ct,' ')))
row.names(ctparts) = NULL
library(chron)
timecrt = chron(dates = ctparts[,1],times = ctparts[,2],format = c('d/m/y','h:m:s'))
learn <- data.frame(learn,timecrt)
learn$time<-NULL
learn$daycrt <- dates(learn$timecrt) ## extract the day an item was created
learn <- learn[order(learn$userid,learn$timecrt),]
```

```
#count of items per user to assist in the REPEAT loop belo
ci.userid <- data.frame(aggregate(learn$item,by = list(learn$userid),FUN = "count"))
colnames(ci.userid) <- c("userid","citem")
ci.userid <- ci.userid[order(ci.userid$userid),]

tlen <- rep(NA,nrow(learn)) # time length between successive time points
wtime <- rep(NA,nrow(learn)) # time variable within all study sessions
wt1 <- rep(NA,nrow(learn)) # time variable within specific study sessions
btime <- rep(NA,nrow(learn)) # time variable between all study sessions
ses <- rep(NA,nrow(learn)) # session count
bstime <- rep(0,nrow(learn)) # time variable between specific study sessions
i<-1 # counter
c <-1 # counter
len <- ci.userid$citem[c]

repeat{
tlen[i]<- 0
ses[i] <- 1
bstime[i] <- tlen[i]
wt1[i] <- wtime[i]<- tlen[i]
btime[i]<- bstime[i]
b <- i+1
for (i in b:len){
   tlen[i]<-(learn$timecrt[i]-learn$timecrt[i-1])*24
      if (learn$daycrt[i] = = learn$daycrt[i-1]){ses[i]<-ses[i-1]
         bstime[i] <- bstime[i-1]
         wtime[i] <- tlen[i]+wtime[i-1]
         btime[i] <- btime[i-1]
         wt1[i] <- tlen[i]+wt1[i-1]
}
      else {ses[i]<-1+ses[i-1]
         bstime[i] <- tlen[i]
         wtime[i] <- tlen[i]+wtime[i-1]-bstime[i]
         btime[i] <- btime[i-1]+bstime[i]
         wt1[i] <- tlen[i]-bstime[i]
}
}
i <- i+1
c <-c+1
len <- i+ci.userid$citem[c]-1
if (c>nrow(ci.userid)){break}
}
learn <- data.frame(learn,tlen,bstime,ses,wtime,btime,wt1)
learn <- learn[order(learn$userid,learn$timecrt),]
learn$timecrt <- NULL; learn$daycrt <- NULL # drop unwanted variables

#### Retain those persons with > = 2 study sessions #####
library(sqldf)
cses.userid <- sqldf("select distinct(userid),count(distinct(ses)) as cses from learn group by userid")
learn <- merge(learn,cses.userid,by = "userid")
learn <- subset(learn, cses%in% cses[which(cses>1)])
learn$cses <- NULL
##Change between session time to days
```

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

```
learn$bstime <- (learn$bstime)/24
learn$btime <- (learn$btime)/24

##START#############################################
## Prepare Data for use in WinBUGS###
##Make dummies for sessions for use in WinBUGS ##
for (i in 1: nrow(learn)){
   if (learn$ses[i] = = 1){learn$ses1[i]<-1}
      else{learn$ses1[i] <- 0}
   if (learn$ses[i] = = 2){learn$ses2[i]<-1}
      else{learn$ses2[i] <- 0}
   if (learn$ses[i] = = 3){learn$ses3[i]<-1}
      else{learn$ses3[i] <- 0}
}
## Transfer Data for analysis in WinBUGS###
iduser<- cbind(sqldf("select distinct(userid) from learn"),iduser = c(1:length(unique(learn$userid))))
iditem<- cbind(sqldf("select distinct(item) from learn"),iditem = c(1:length(unique(learn$item))))
idall<- merge(iduser,iditem)
learn <- merge(idall, learn, by = c("userid","item"), all = FALSE)
rm(iduser,iditem,idall)
sdat1 <- subset(learn, select = c(iduser,iditem,grade,wt1,ses1,ses2,ses3))
write.table(sdat1,"sdat1.txt",append = F,row.names = F, col.names = T)
rm(sdat1)
##END###############################################

## Model fitting continued in R#
library(lme4)
learn$ses <- as.factor(learn$ses)
### Fit Model 3
M3.1 <- lmer(grade~-1+ses+ses:wt1+ses:I(wt1^2)+(-1+ses+ses:wt1+ses:I(wt1^2)|userid)+(1|item),
data = learn, family = binomial(link = "logit"))
M3.2 <-
lmer(grade~-1+ses+ses:wt1+(-1+ses+ses:wt1|userid)+(1|item),data = learn, family = binomial(link =
"logit"))
## Fit Model 4
M4.1 <- lmer(grade~1+wtime+I(wtime^2)+btime+I(btime^2)+(1+wtime+I(wtime^2)+btime+I
(btime^2)|userid)+(1|item),data = learn, family = binomial(link = "logit"))
M4.2 <- lmer(grade~1+wtime+btime+(1+wtime+btime|userid)+(1|item),data = learn, family = bino-
mial(link = "logit"))
##Scatter matrix plot for random effects' estimates
library(car)
scatterplotMatrix(ranef(M4.2)$userid, var.labels = c("intercept","wtime","btime"),cex.label = 1.3,diagonal
= c("histogram"),plot.points = TRUE)
## Fit Model 6
M6.2 <- lmer(grade~-1+course+course:wtime+course:btime+(-1+course+course:wtime+course:
btime|userid)+(1|item),data = learn, family = binomial(link = "logit"))
```

WinBUGS Macro

```
## EQ3: Session-Specific Growth IRT model for statistics online data ##
model {
# M observations
for (i in 1:M) {
```

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

```
      grade[i] ~ dbern(p[i])
      logit(p[i]) <- b01*ses1[i] + b02*ses2[i] + b03*ses3[i] +
      b11*ses1[i]*wt1[i] + b12*ses2[i]*wt1[i] + b13*ses3[i]*wt1[i] +
      si1*u[userid[i]]*ses1[i] + si2*u[userid[i]]*ses2[i]+
      si3*u[userid[i]]*ses3[i] + sl1*u[userid[i]]*ses1[i]*wt1[i] +
      sl2*u[userid[i]]*ses2[i]*wt1[i]+ sl3*u[userid[i]]*ses3[i]*wt1[i] + v[itemid[i]]
}
# N Persons
for (n in 1:N) {
      u[n] ~ dnorm(0,1)}
# J Items
for (j in 1:J) {
      v[j] ~ dnorm(0,taui)}
# Priors
b01~ dnorm(0.0,1.0E-6)
b02~ dnorm(0.0,1.0E-6)
b03~ dnorm(0.0,1.0E-6)
b11~ dnorm(0.0,1.0E-6)
b12~ dnorm(0.0,1.0E-6)
b13~ dnorm(0.0,1.0E-6)

# Hyperpriors
si1 ~ dnorm(0,1)I(0,)
si2 ~ dnorm(0,1)I(0,)
si3 ~ dnorm(0,1)I(0,)
sl1 ~ dnorm(0,1)I(0,)
sl2 ~ dnorm(0,1)I(0,)
sl3 ~ dnorm(0,1)I(0,)
taui ~ dgamma(1,1)

# variance of random intercepts
vsi1 <- pow(si1,2)
vsi2 <- pow(si2,2)
vsi3 <- pow(si3,2)
# variance of random slopes

vsl1 <- pow(sl1,2)
vsl2 <- pow(sl2,2)
vsl3 <- pow(sl3,2)
# variance of items
vbeta <- 1/sqrt(taui)
}
list(M = 3073,N = 64,J = 145) ### load data

userid[] itemid[] grade[] wt1[] ses1[] ses2[] ses3[]
1 78 1 2.36 0 1 0
1 52 1 2.04 0 1 0
... ... ... ... ... ... ...
... ... ... ... ... ... ...
64 36 1 0.49 1 0 0
64 5 0 4.17 1 0 0
```

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015

END

list(b01 = 0.1,b02 = 0.1,b03 = 0.1,b11 = 0.1,b12 = 0.1,b13 = 0.1,si1 = 1,si2 = 1,si3 = 1,sl1 = 1,sl2 = 1,sl3 = 1,taui = 1)## initial values

EQ4: Growth IRT model for statistics online data

```
model {
# M observations
for (i in 1:M) {
      grade[i] ~ dbern(p[i])
      logit(p[i]) <- b0 +b1*wtime[i] + b2*btime[i]+ u[userid[i]] +
      slw*u[userid[i]]*wtime[i] + slb*u[userid[i]]*btime[i]+ v[itemid[i]]
}
# N Persons
for (n in 1:N) {
      u[n] ~ dnorm(0,taup)}
# J Items
      for (j in 1:J) {
      v[j] ~ dnorm(0,taui)}
# Priors
b0~ dnorm(0.0,1.0E-6)
b1~ dnorm(0.0,1.0E-6)
b2~ dnorm(0.0,1.0E-6)

# Hyperpriors
taui ~ dgamma(1,1)
taup ~ dgamma(1,1)
vtheta <- 1/sqrt(taup)
vbeta <- 1/sqrt(taui)
slw ~ dnorm(0,1)I(0,)
slb ~ dnorm(0,1)I(0,)
vslw <- pow(slw,2)
vslb <- pow(slb,2)
}
list(M = 3235,N = 64,J = 145) ### load data
“iduser” “iditem” “grade” “wtime” “btime1”
1 78 1 2.35 0.09
1 52 1 2.04 0.09
1 65 1 2.23 0.09
... ... ... ... ... ... ...
... ... ... ... ... ... ...
64 36 1 0.49 0
64 5 0 4.11 0
END
```

list(b0 = 0.1,b1 = 0.1,b2 = 0.1,taup = 1,taui = 1,slw = 1,slb = 1)## initial values

Downloaded by [KU Leuven University Library] at 11:07 10 February 2015