An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments

Park, J.Y.; Joo, S.-H.; Cornillie, F.; van der Maas, H.L.J.; Van den Noortgate, W.

DOI
10.3758/s13428-018-1166-9

Publication date
2019

Document Version
Final published version

Published in
Behavior Research Methods

Behavior Research Methods (2019) 51:895–909
https://doi.org/10.3758/s13428-018-1166-9

An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments

Jung Yeon Park[1] · Seang-Hwane Joo[1] · Frederik Cornillie[1] · Han L. J. van der Maas[2] · Wim Van den Noortgate[1]

Published online: 3 December 2018

Abstract
Electronic learning systems have received increasing attention because they are easily accessible to many students and are capable of personalizing the learning environment in response to students' learning needs. To that end, using fast and flexible algorithms that keep track of the students' ability change in real time is desirable. Recently, the Elo rating system (ERS) has been applied and studied in both research and practical settings (Brinkhuis & Maris, 2009; Klinkenberg, Straatemeier, & van der Maas in *Computers & Education, 57,* 1813–1824, 2011). However, such adaptive algorithms face the cold-start problem, defined as the problem that the system does not know a new student's ability level at the beginning of the learning stage. The cold-start problem may also occur when a student leaves the e-learning system for a while and returns (i.e., a between-session period). Because external effects could influence the student's ability level during the period, there is again much uncertainty about ability level. To address these practical concerns, in this study we propose alternative approaches to cold-start issues in the context of the e-learning environment. Particularly, we propose making the ERS more efficient by using an explanatory item response theory modeling to estimate students' ability levels on the basis of their background information and past trajectories of learning. A simulation study was conducted under various conditions, and the results showed that the proposed approach substantially reduces ability estimation errors. We illustrate the approach using real data from a popular learning platform.

Keywords E-learning system · Cold-start problem · Elo rating system · Explanatory IRT · Between-session effect

There has been an increasing trend toward implementing electronic learning (e-learning) environments for higher education as well as for K–12 education, because advanced technologies can have substantial advantages in assisting students' learning. One important advantage of technology-based learning environments is accessibility, with students having access to the learning environment at their own pace, anytime and anyplace. In addition, e-learning environments can be more effective and efficient than traditional classroom learning, because they are capable of personalizing students' learning opportunities by using an adaptive system (Brusilovsky and Peylo, 2003). Unlike static learning environments, where the same contents and information are given to each student, the adaptive systems in e-learning environments can take students' individual characteristics into account. For example, students' learning preferences (e.g., visual, auditory, or kinesthetic), background information (e.g., gender, age, and socio-economic status), and knowledge level (e.g., previous courses taken and education level) can be used as information that the adaptive system incorporates so as to optimize the learning conditions. The importance of accounting for individual characteristics in e-learning environments has consistently been emphasized in previous studies (e.g., Kalyuga & Sweller, 2005; Shute & Towle, 2003; Snow, 1989, 1996).

One of the systems used in adaptive e-learning environments is the Elo rating system (ERS; Elo, 1978). The ERS can be used to produce adaptive item sequencing, in which items are selected in real time on the basis of the current estimate of the student's ability or knowledge level (Wauters, Desmet, & Van den Noortgate, 2010). More specifically, when a student responds to an item, the ERS algorithm computes updated estimates of the student's ability level and the item difficulty, based on the correctness of the current response. Then the next item is provided, such that the item difficulty level matches the student's current ability level. The ERS not only can be used to estimate a constant student's ability level, but is also applicable to tracking

✉ Jung Yeon Park
ellie.park@kuleuven.be

[1] Faculty of Psychology and Educational Sciences, imec–ITEC, KU Leuven, Leuven, Belgium

[2] Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

a changing ability level in learning environments where the student's ability level is expected to improve, because of the instant feedback that is provided on each response (Wauters et al., 2010). Note that although the ERS can be used to gradually obtain reliable estimates of both a student's abilities and the item difficulties, adaptive item sequencing would be more efficient if we could start from a precalibrated item bank, including information on item difficulty and possibly on other characteristics of the items. In previous research, the development and application of the ERS has been widely explored (e.g., Brinkhuis & Maris, 2009; Coomans, Hofman, Brinkhuis, van der Maas, & Maris, 2016; Klinkenberg, Straatemeier, & van der Maas, 2011; Maris & van der Maas, 2012; Savi, van der Maas, & Maris, 2015).

When a new student comes into the e-learning system, the ERS algorithm is required to set an initial value for the student's ability. Initial values may be drawn randomly, or the mean of the previous students' ability levels can be used (Wauters et al., 2010). However, an ambiguous initial value could lead to inaccurate ability estimation, resulting in a higher number of responses being required in order to get accurate ability estimates, or in higher standard errors of the ability estimates given a particular number of responses (Wauters et al., 2010). As a consequence, it may take longer before the environment is optimally adapted to the student. This problematic situation is referred as the *cold-start problem*.

In general, the cold-start problem occurs when a student starts working in the learning environment. However, the cold-start problem may also occur if a student decides to leave the e-learning system for a while before starting a new session. Between-session periods are prevalent in the e-learning environment because the students have flexibility in when to access this environment. There may be ability change in relation to various experiences during between-session periods. For example, the student might take extra training through printed learning materials or additional online learning platforms. For such cases, the student's ability tends to increase. Alternatively, the student might not have been involved in any type of constructive learning, and might have partially forgotten the relevant content. In this case, his or her ability tends to decrease. For that reason, roughly assuming that the ability estimated from the previous session will be the same as the ability level at the beginning of the next session, by ignoring potential ability change between study sessions, may again lead to inefficient ability estimates. The longer the between-session period, therefore, the higher the uncertainty about the student's ability at the start of the new session, and therefore the more we can consider the start of a new session a cold start.

One possible solution to both cold-start problems would be to incorporate additional information about the student. More specifically, a student's background information may give a better idea of the initial ability level and how the ability level may have changed between sessions. However, the current ERS uses the Rasch model, which does not allow the flexibility to utilize the student's information, because the model includes only two parameters (e.g., item difficulty and latent ability). Alternatively, explanatory item response theory (explanatory IRT; De Boeck & Wilson, 2004) analyses can be used to explore the effects of background variables on initial abilities and evolution in ability. Explanatory IRT models have been investigated for educational data in which student and item effects on the probability of a correct response have been considered random (Van den Noortgate, De Boeck, & Meulders, 2003), allowing the inclusion of student and item characteristics as predictors. Furthermore, data from an e-learning environment were analyzed with the explanatory IRT model in another study (Kadengye, Ceulemans & Van den Noortgate, 2014). However, the explanatory IRT model has not yet been used in combination with the ERS algorithm. Therefore, it is unknown to what extent the cold-start problems may be resolved by adapting explanatory IRT with the ERS. Thus, in the present study we aimed to investigate the cold-start problem in the e-learning system, both when new students come in to the learning environment or students come back to the environment after a specific period between sessions. We further aimed to describe and empirically evaluate one possible approach to addressing the cold-start problem, by combining explanatory IRT and the ERS.

Below, we first give more details about the ERS and its proposed improvement by means of explanatory IRT. Next, we describe a simulation study evaluating the performance of this combined approach. We then demonstrate its applicability to real-life data collected from an e-learning platform.

The Elo rating system

The ERS was originally developed as an adaptive estimation method for chess players' ability. Adapting the concept of a chess game to the educational measurement, a student is considered a player, an item is considered an opponent, and a correctly answered item is considered a win for the student. In an adaptive e-learning environment, the ERS allows testing to keep track of individualized learning growth as soon as students have engaged in a sequence of items. The ERS estimates the individual's trajectory by successively updating the ability estimates. To illustrate the ERS process, consider $\theta_{p(t)}$ to be the ability of a student p after solving an item at measurement occasion t. Also, suppose $Y_{pi(t)}$ to be the outcome for student p on item i measured at measurement occasion t, where the outcome is dichotomously scored (0 = *incorrect answer* and 1 = *correct answer* to the item). Then, the ERS for updating the ability parameter takes the following sequence:

$$\hat{\theta}_{p(t)} = \hat{\theta}_{p(t-1)} + K\left\{Y_{pi(t)} - E\left(Y_{pi(t)}\right)\right\}, \quad (1)$$

where $\hat{\theta}_{p(t-1)}$ is the ability estimate at the previous measurement occasion $t-1$ for student p, and $E(Y_{pi(t)})$ is the expected response for the current measurement occasion t. K is a step size, which provides a weight of the difference between the current and expected responses for student p. Several studies have explored the optimal step size K for the ERS (e.g., Glickman, 1999; Klinkenberg et al., 2011; Papousek, Pelánek, & Stanislav, 2014), and $K = 0.4$ has typically been used in the context of learning analytics (Wauters, Desmet, & Van den Noortgate, 2012). The expected response $E(Y_{pi(t)})$ for person p on item i and measurement occasion t—with $Y_{pi(t)} = 1$ if the answer was correct, and $Y_{pi(t)} = 0$ otherwise—can be computed by using the Rasch model, more specifically as a function of the difference between the ability as estimated before the response was given, $\hat{\theta}_{p(t-1)}$, and the difficulty of the item, β_i:

$$logit\left(E\left[Y_{pi(t)}\right]\right) = logit\left(P\left[Y_{pi(t)} = 1\right]\right) = \hat{\theta}_{p(t-1)} - \beta_i. \quad (2)$$

Note that in Eq. 1, it is necessary to specify a starting value for the ability, $\hat{\theta}_{p(0)}$, in order to initiate the ERS algorithm. In principle, zero (or a randomly drawn value from, e.g., a standard normal distribution) can be used, because the mean ability is often assumed to be zero in order to make the Rasch model identified. However, to avoid the cold-start problem, several alternative methods to specify the initial value for an adaptive-learning system have been suggested (e.g., Bobadilla, Ortega, Hernando, & Bernal, 2012; Pereira & Hruschka, 2015; Tang & McCalla, 2004). Wauters et al. (2010) suggested an individual-specific approach for an adaptive e-learning environment that uses explanatory IRT to obtain more specific starting values using students' information.

Explanatory IRT

The explanatory IRT is a type of multilevel model in which students' item responses are considered as the first-level observations, students are considered as second-level units, and the students' and/or items' characteristics are included as predictors (De Boeck & Wilson, 2004). Suppose that there a total of J student background variables are available; the mathematical expression of the explanatory IRT model can then be described as follows:

$$logit\left(P\left[Y_{pi(t)} = 1\right]\right) = \alpha_0 + \sum_{j=1}^{J} \alpha_j Z_{pj} + w_p - \beta_i \quad (3)$$

where $P[Y_{pi(t)} = 1]$ is the probability of correctly answering item i by student p at measurement occasion t, α_0 is the intercept, Z_{pj} is the score of person p on the jth explanatory variable, α_j is the effect of the jth student characteristic, and w is the random deviation of the students after the effects of student information are taken into account. Finally, β_i is the difficulty level of item i. Note that the item difficulty β_i is assumed to be a random factor that follows a normal distribution with a zero mean and a variance of σ^2_β. An individual-specific ability estimate for student p then can be obtained from the formula $\theta_{p(t)} = \alpha_0 + \sum_{j=1}^{J} \alpha_j Z_{pj} + w_p$.

To account for changing ability, we have to model the ability of the person as a function of time. In addition, if there are between-session periods in the e-learning environment, the ability trajectory in the between-session periods should also be modeled, in order to predict the ability of a new student at the start of a new session. To do so, we can create two distinct variables, following Kadengye, Ceulemans, and Van den Noortgate (2014): $wtime_{p(t)}$ and $btime_{p(t)}$. The first, $wtime_{p(t)}$, is a time variable that records the accumulated time within sessions until student p solves an item at measurement occasion t; $btime_{p(t)}$ is a time variable that records the accumulated time for the between-session periods (i.e., the period during which the student is not engaged in the learning environment) for a student p until occasion t. During each session, $wtime_{p(t)}$ continuously increases, while $btime_{p(t)}$ remains constant. Similarly, between sessions, $btime_p$ increases, while $wtime_p$ remains constant. By specifying $wtime_p$ and $btime_p$ in this way, their coefficients refer to ability growth (i.e., the slope effect) within and between sessions, respectively. To illustrate the structure of the data set more clearly, Table 1 presents the variables for the explanatory IRT model. This table includes an example of a student's study time and assumes three study sessions with two between-session periods (24 h each). For Study Sessions 1, 2, and 3, the student solves totals of n_{11}, n_{12}, and n_{13} items, respectively, and each session took t_1, t_2, and t_3 hours to finish. For simplicity, we assume that the time used for each item is 0.1 h, or 6 min, but of course in reality the time required can vary over items.

To incorporate ability changes in function of $wtime_{p(t)}$ and $btime_{p(t)}$, we can extend Eq. 3 to the following explanatory IRT model:

$$\begin{aligned} logit\left(P\left[Y_{pi(t)} = 1\right]\right) = &\left(\alpha_{00} + \sum_{j=1}^{J} \alpha_{0j} Z_{pj} + w_{0p}\right) + \left(\alpha_{10} + \sum_{j=1}^{J} \alpha_{1j} Z_{pj} + w_{1p}\right) wtime_{p(t)} \\ &+ \left(\alpha_{20} + \sum_{j=1}^{J} \alpha_{2j} Z_{pj} + w_{2p}\right) btime_{p(t)} - \beta_i, \end{aligned} \quad (4)$$

where the first component, $\left(\alpha_{00} + \sum_{j=1}^{J} \alpha_{0j} Z_{pj} + w_{0p}\right)$, is the ability at the initial measurement occasion $t = 0$ (i.e., when both time variables are equal to zero). The second component in Eq. 4, $\left(\alpha_{10} + \sum_{j=1}^{J} \alpha_{1j} Z_{pj} + w_{1p}\right)$, represents the ability change over time within sessions. Similarly, the third component in Eq. 4, $\left(\alpha_{20} + \sum_{j=1}^{J} \alpha_{2j} Z_{pj} + w_{2p}\right)$, represents the ability change between sessions. In the equation, α_{00}, α_{10}, and α_{20} refer, respectively, to the expected initial ability, ability change within sessions, and ability change between sessions when the student predictor variables are equal to zero. Similarly, α_{0j}, α_{1j}, and α_{2j} are the effects of student variable j on the initial

Table 1 Structure of a data set

Student	Session	Item	Score	btime	wtime	Time
1	1	1	1	0	0.1	0.1
1	1	2	0	0	0.2	0.2
1	1	3	1	0	0.3	0.3
1	1	.	.	0	.	.
1	1	.	.	0	.	.
1	1	n_{11}	0	0	t_1	t_1
1	2	$n_{11}+1$	0	24	0.1	$t_1+24+0.1$
1	2	$n_{11}+2$	0	24	0.2	$t_1+24+0.2$
1	2	$n_{11}+3$	1	24	0.3	$t_1+24+0.3$
1	2	.	.	24	.	.
1	2	.	.	24	.	.
1	2	$n_{11}+n_{12}$	1	24	t_2	t_1+24+t_2
1	3	$n_{11}+n_{12}+1$	1	48	0.1	$t_1+t_2+48+0.1$
1	3	$n_{11}+n_{12}+2$	0	48	0.2	$t_1+t_2+48+0.2$
1	3	$n_{11}+n_{12}+3$	1	48	0.3	$t_1+t_2+48+0.3$
1	3	.	.	48	.	.
1	3	.	.	48	.	.
1	3	$n_{11}+n_{12}+n_{13}$	1	48	t_3	$t_1+t_2+48+t_3$

ability, the ability change within sessions, and the ability change between sessions, respectively. Finally, w_{0p}, w_{1p}, and w_{2p} are the random deviations of the student p estimates after the students' information is taken into account. Given the student information for a new student, as well as estimates of the coefficients of Eq. 4, the predicted ability of student p at various time points therefore can be obtained as follows: $\hat{\theta}_{p(t)} = (\hat{\alpha}_{00} + \sum_{j=1}^{J}\hat{\alpha}_{0j}Z_{pj}) + \left(\hat{\alpha}_{10} + \sum_{j=1}^{J}\hat{\alpha}_{1j}Z_{pj}\right)wtime_{p(t)} + (\hat{\alpha}_{20} + \sum_{j=1}^{J}\hat{\alpha}_{2j}Z_{pj})btime_{p(t)}$. For a student who has already answered several items, we can make even better predictions, by estimating and accounting for the student-specific random effects w_{0p}, w_{1p}, and w_{2p}, as well.

Model estimation

The parameters of the explanatory IRT model (including the random effects w_{0p}, w_{1p}, and w_{2p}) shown in Eq. 4, can then be estimated within the Bayesian framework (e.g., Dai & Mislevy, 2009; Frederickx, Tuerlinckx, De Boeck, & Magis, 2010; Kadengye et al., 2014). Bayesian inference draws conclusions about parameters in the form of a posterior distribution that combines the likelihood of the data with prior knowledge about the parameters. Let (Y, $\boldsymbol{\Omega}$) denote the likelihood function of a set of all parameters shown in Eq. 4, say $\boldsymbol{\Omega}$, and let $P(\boldsymbol{\Omega})$ denote the prior distribution of those parameters. The posterior distribution is proportional to the product of the two components, $P(\boldsymbol{\Omega}|\,\boldsymbol{Y}) \propto L(\boldsymbol{Y},\boldsymbol{\Omega})\ P(\boldsymbol{\Omega})$. Because obtaining an analytical solution for $P(\boldsymbol{\Omega}|\,\boldsymbol{Y})$ may not be feasible or may be extremely intensive to compute, a Markov chain Monte Carlo (MCMC) algorithm is typically employed.

Table 2 Parameter estimates of the explanatory IRT models (n = 250 training set)

	$\sigma_{w2}=0.5$			$\sigma_{w2}=1.5$		
Parameter	True	Est.	*SE*	True	Est.	*SE*
Intercept						
Nondyslexia	0.1	0.135	0.093	1.5	1.175	0.123
Dyslexia	− 2.5	− 2.035	0.166	− 2.5	− 2.159	0.182

Initial ability of new students As a first way to enhance the traditional ERS, an explanatory IRT analysis can be used to give more accurate starting values when the ERS has started updating the ability estimates of new student(s). To estimate the initial ability of the new students, we fit the model to data that have already been collected from the previous students in the environment. Given the new students' background information, this allows us to obtain an estimated initial ability, $\left(\hat{\alpha}_{00} + \sum_{j=1}^{J}\hat{\alpha}_{0j}Z_{pj}\right)$. In the case that there is no prior knowledge about α_{00} and α_{0j}, it is typical to choose a noninformative or vague prior distribution for the Bayesian inference (Gelman et al., 2014). Therefore, in this study a normal distribution with an extremely large variance (i.e., small precision) was chosen for the prior distribution of initial ability, α_{00} and α_{0j}~$N(0, 10^6)$.

Ability change between study sessions As a second means to enhance the performance of the traditional ERS, the explanatory IRT analysis can give information about ability change between sessions (i.e., the effect of $btime_{p(t)}$ in Eq. 4). The information assists in choosing more accurate starting values when the traditional ERS has restarted updating the ability for the next session. Similar to the estimation of the initial ability (i.e., $\hat{\alpha}_{00} + \sum_{j=1}^{J}\hat{\alpha}_{0j}Z_{pj}$), the estimate of a slope parameter $\left(\hat{\alpha}_{20} + \sum_{j=1}^{J}\hat{\alpha}_{2j}Z_{pj}\right)$ for $btime_{p(t)}$ in Eq. 4 provides the ability change that can be expected for a student with these characteristics. Adding this estimated change to the ability estimate at the end of the previous session gives the starting value for resuming the ERS at the beginning of the next session. If for a given student we already have observed data for at least two sessions, the explanatory IRT model can also estimate w_{2p} from Eq. 4 for the particular student, on top of the expected

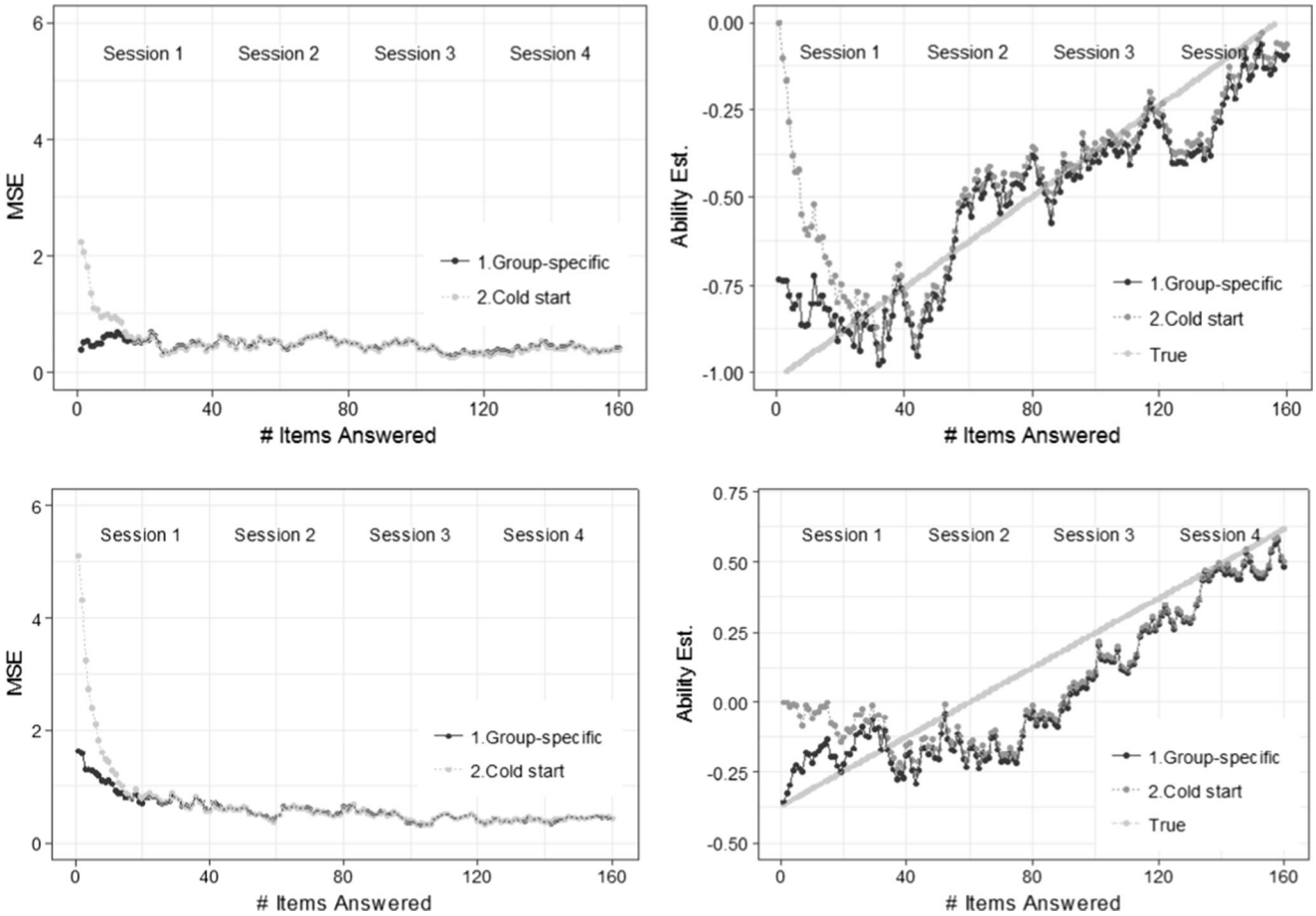

Fig. 1 *MSE*s (left) and true versus estimated ability trajectories (right) for an average student. Note that the upper panels are for Scenario 1 and the lower panels are for Scenario 2 in Study 1

ability change corresponding to the specific student characteristics.

In a real-life setting, estimations of fixed and random effects can be updated by fitting explanatory IRT to the data consecutively collected from further sessions. Suppose that the posterior distribution of α_{2j} and w_{2p} after the current session can be formulated by

$$\begin{aligned} P(\alpha_{2j}|Y) &\propto L(Y_p, \alpha_{2j})\, P(\alpha_{2j}), \\ P(w_{2p}|Y) &\propto L(Y_p, w_{2p})\, P(w_{2p}). \end{aligned} \tag{5}$$

In this case, the prior distributions $P(\alpha_{2j})$ and $P(w_{2p})$ based on data up to the current session can be informed by the posterior distribution based on the data up to the previous session. Therefore, we assign our prior for the current session such that $\alpha_{2j} \sim N(\mu_{\alpha 2}, \sigma^2_{\alpha 2})$, where $\mu_{\alpha 2}$ and $\sigma^2_{\alpha 2}$ are informed by the posterior mean and variance based on the previous session. Similarly, the prior for $\hat{w}_{2p}$ also is to be updated along with $\hat{\alpha}_{2j}$ across the continuing analysis after each session. That is, $w_{2p} \sim N(\mu_{w2}, \sigma^2_{w2})$, where μ_{w2} and σ^2_{w2} are informed by the posterior mean and variance based on the previous session. Allowing the normal prior density with an informative mean and variance for w_{2p} is advantageous for obtaining its posterior mean, $\hat{w}_{2p}$, especially when the students have not gone through enough sessions. With only a few observations, the posterior inference about ability change, w_{2p}, is shifted away from the individual-specific estimate toward the direction of the population-averaged estimate. With more observations after more sessions, the inference relies more on individual-specific change. Therefore, this characteristic enables the analysis to adopt an average student's slope estimate plus the student-specific deviation, which attempts to approximates more "individualized" ability change between sessions. Note that it would be natural to assign noninformative priors for the very first session.

Simulation study

To provide evidence of the performance of the proposed method, a simulation study was conducted. We produce simulated data using a data generation process that mimics

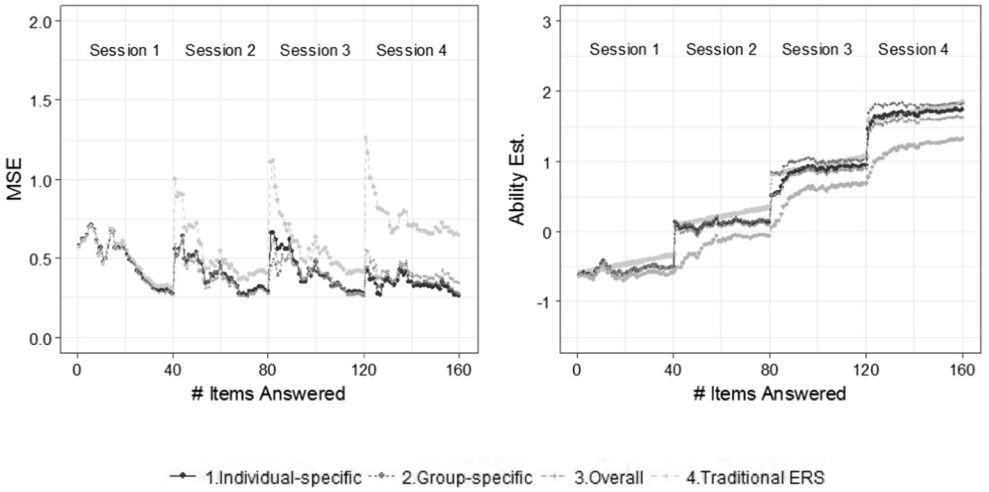

Fig. 2 *MSE*s (left) and estimated ability trajectories for an average student (right; straight lines indicate the true ability trajectory averaged for the group) when there is constructive learning with small student variation during the period between sessions. Note that with the validation data (n = 50 new students), the initial ability estimates in the ERS for the students were determined by group-specific starting values

various aspects of an online learning environment. In accordance with our research questions, the full simulation study consisted of two parts. In Study 1, we examined the effects of addressing the cold-start problem for new students. Here we assumed that students' true ability evolves while they are being engaged in the learning environment, but that there are no ability changes while they are outside the environment (i.e., between sessions). In Study 2, we focused on exploring the effects of addressing ability changes between sessions. Therefore, ability growth both within and between sessions would be simulated.

For each condition of the two studies, we simulated data for 300 students who engaged in a total of four sessions and solved 40 items per session. In particular, the simulated data sets contained the students' responses to a total of 160 items, which were randomly assigned out of an item bank of 1,000 items. Then each data set was divided into two subsets: (a) a "training set" (n = 250 students) was used to fit the explanatory IRT model, in order to obtain the starting values in the ERS, and (b) a "validation set" (n = 50 students) was used to execute the ERS by incorporating the information from the training set. Individuals from the validation set were considered new students who had just engaged in the environment. For those students, we allowed the ERS to estimate their ability growth trajectory within sessions. Note that evaluation of an ERS with the support of explanatory IRT modeling was the primary purpose of the present study, and thus we compared the performance of the proposed ERS with a traditional ERS in which no student information was taken into account. In both subsets, we assumed that the data were collected from two groups (i.e., nondyslexic and dyslexic groups) with equal sample sizes (i.e., a balanced design).

Study 1: No true ability change between sessions

Study design

To generate the true ability trajectory of the students, the following explanatory IRT model was used:

$$logit\left(P\left[Y_{pi(t)}=1\right]\right)=\left(\alpha_{01}Z_{p1}+\alpha_{02}Z_{p2}+w_{0p}\right)+\left(\alpha_{11}Z_{p1}+\alpha_{12}Z_{p2}+w_{1p}\right)wtime_{p(t)}-\beta_i \quad (6)$$

Z_{p1} and Z_{p2} are binary indicators for two groups, say a nondyslexic and a dyslexic group, respectively (Z_{p1} = 1 if the student is from the nondyslexic group, and 0 otherwise; Z_{p2} = 1 if the student is from the dyslexic group, and 0 otherwise). The model does not include an intercept, so α_{01} and α_{02} refer to the expected initial abilities in these two groups. The initial-ability parameter is affected not only by the group membership, but also by individual-specific factors (w_{0p}, or the individual deviation from the group mean). Similarly, the learning trend varies over persons around a group-specific mean. We assumed that the nondyslexic group would typically show a greater average initial ability level (α_{01}) and learning growth (α_{02}) than the dyslexic group, since many educators and learning platforms are concerned with the lower learning capability of those who experience dyslexia (e.g., Humphrey & Mullins, 2002; Polychroni, Koukoura, & Anagnostou, 2006; Vukovic, Lesaux, & Siegel, 2010).

Initial ability With regard to the true fixed and random effects for the initial ability levels, two scenarios were considered. In the first scenario, a relatively smaller difference between the two group means and a relatively

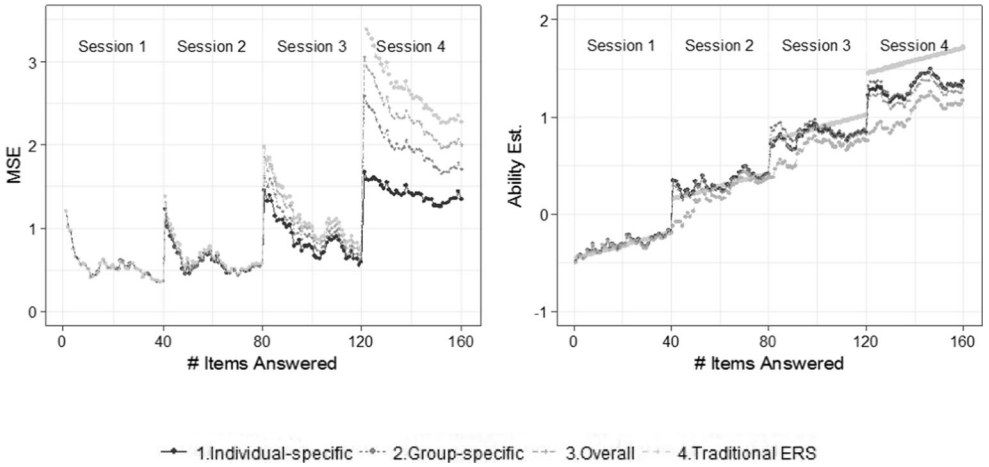

Fig. 3 *MSE*s (left) and estimated ability trajectories for an average student (right; straight lines indicate the true ability trajectory averaged for the group) when there is constructive learning with large student variation during the period between sessions. Note that with the validation data (n = 50 new students), the initial ability estimates in the ERS for the students were determined by group-specific starting values

smaller random standard deviation were considered. In particular, the fixed effects were set at α_{01} = 0.1 and α_{02} = − 2.5, respectively, and the standard deviation of the random effects was set at σ_{w0} = 0.5, where w_{0p}~$N(0, \sigma^2_{w0})$. In the second scenario, the fixed effects were set at α_{01} = 0.5 and α_{02} = − 2.5, respectively, and the standard deviation of the random effects was set at σ_{w0} = 1.5.

Ability growth In both scenarios, the within-session time (i.e., $wtime_{p(t)}$) was simulated to increase by 0.05 h for each item. Also, the fixed time effects were set at α_{11} = 0.18 and α_{12} = 0.05, respectively, and the standard deviation of the random effects was set at σ_{w1} = 0.0008, where w_{1p}~$N(0, \sigma^2_{w1})$.

Item difficulty The true item difficulties $\beta_{p(t)}$, for measurement occasion t for student p, were drawn randomly from $N(0, 2)$. On the basis of those parameter values, the item response of each student (correct/incorrect) for each measurement occasion t was randomly generated on the basis of a binomial distribution with the probability in Eq. 6.

Starting value To initiate the ERS tracing his or her growth trajectory with the validation set (n = 50 students), certain starting values should be given for a new student's initial ability estimates. Two options were considered in this study. First, as is the case with many testing or learning environments, no information was given about the new students, and therefore zero values were assigned for all of them. This is the cold-start situation. For the second option, the explanatory IRT model was fitted to the training set, leading to specific ability levels for the dyslexic or nondyslexic groups. The mathematical formula for this data-analyzing model is as follows:

$$logit\left(P\left[Y_{pi(t)}=1\right]\right)=\left(\sum_{g=1}^{2}\alpha_{0g}Z_{pg}+w_{0p}\right)+\left(\sum_{g=1}^{2}\alpha_{1g}Z_{pg}+w_{1p}\right)wtime_{p(t)}+\left(\sum_{g=1}^{2}\alpha_{2g}Z_{pg}+w_{2p}\right)btime_{p(t)}-\beta_i \quad (7)$$

Implementation For the estimation procedure with the training set, the MCMC algorithm was implemented with R 3.3.3 (R Core Team, 2013). More specifically, JAGS (Plummer, 2015) was implemented in the R package R2jags (Su & Yajima, 2015), which provides wrapper functions for the Bayesian analysis program. For each analysis with JAGS, four chains were run, and each ran for 10,000 iterations. We use a thinning parameter of four and used the first half as burn-in. The resulting iterations from each chain were pooled and randomly mixed after burn-in to be used as samples from the posterior distribution of the parameters. The use of multiple chains and thinning served to reduce the dependencies among the iterations and ensure adequate convergence to and mixing from the posterior distribution. Gelman and Rubin's (1995) statistics were used as convergence diagnostics. The final estimates of the model parameters were obtained by taking the mean of the posterior samples after the burn-in periods.

Once the explanatory IRT models had been fitted, the initial ability values (i.e., the starting values) were obtained from the two approaches (taking zero for everybody, or using the group-specific means). Those initial values were then used to initiate the ERS algorithm. The student ability trajectory across measurement occasions was then estimated and the accuracy of the ability estimates from the ERS was summarized by the mean squared error (*MSE*). In particular, the

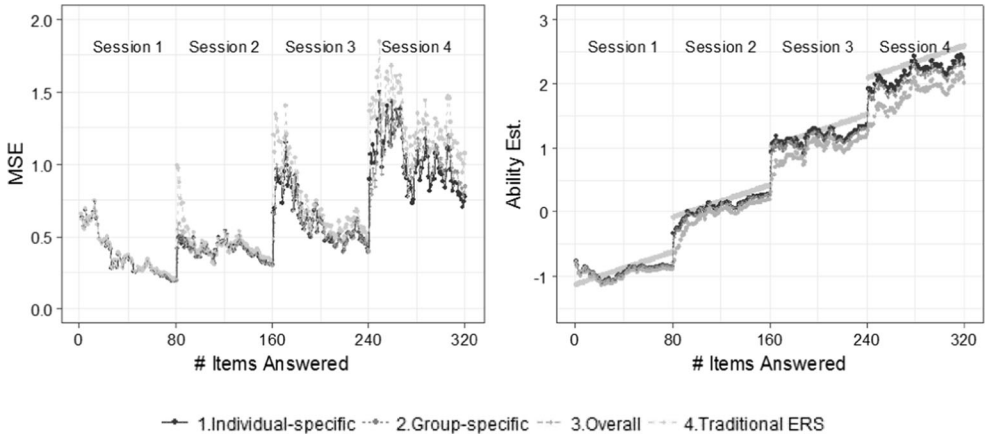

Fig. 4 *MSE*s (left) and estimated ability trajectories for an average student (right; straight lines indicate the true ability trajectory averaged for the group) when there is ability change with small student variation during the period between sessions. Note that with the validation data (n = 50 new students; 80 items per session), the initial ability estimates in the ERS for the students were determined by group-specific starting values

difference between the true and estimated abilities at measurement occasion t was quantified and averaged over the entire sample size of new students—that is,

$$MSE\left(\hat{\theta}_{(t)}\right) = \frac{\sum_{p=1}^{P}\left(\hat{\theta}_{p(t)} - \theta_{p(t)}\right)^2}{P}. \quad (8)$$

Results

Table 2 presents the initial ability levels estimated by the explanatory IRT analysis from the training data (n = 250 students). Specifically, Table 2 shows the posterior means ("Est.") and standard deviations ("*SE*") from the models, fit for two different scenarios. On the basis of the first scenario, the estimated initial ability levels ("intercept") were equal to 0.135 and – 2.035 for the nondyslexic and dyslexic groups, respectively. On the other hand, the second scenario concerned a larger difference between the two groups—that is, 1.175 and – 2.159, respectively.

With the validation data (n = 50 new students), the ERS is initiated by using two options: a cold start (initial ability levels for all new students being at zero) or the group-specific start, based on the explanatory IRT model (under "Intercept" in Table 2). Figure 1 summarizes the results of the two scenarios: Scenario 1 (upper panels) and Scenario 2 (lower panels). In the same figure, the panels in the left column show *MSE*s along with the total number of items answered (at each measurement occasion from the ERS). Also, the panels in the right column illustrate the true and the estimated ability trajectories for an average student.

Each plot in the figure displays the performance when using the group-specific start values as compared to the cold-start values. Specifically, using the group-specific values suggests that the explanatory IRT approach decreases the systematic bias in the initial ability estimates (as can be seen in the right panels). As a result of this decreased bias, the initial *MSE* is also smaller for the explanatory IRT approach (as can be seen in the left panels). However, because students' ability may still deviate from the ability that is expected on the basis of their background characteristics, there is still variation in the ability estimates around the overall value, adding to the *MSE*.

When the group difference in initial ability is small (in the upper left panel of Fig. 1), the *MSE*s begin with 2.4 and 0.5 for the cold-start and the group-specific starts, respectively. When the group difference is large (in the lower left panel of Fig. 1), the *MSE*s begin with approximately 5.2 and 1.8 for the cold-start and group-specific starts, respectively. This suggests that the *MSE*s in the second scenario are generally greater than those in the first scenario, since the simulated data variance was large. For both scenarios, the *MSE* tends to gradually decrease as the number of items answered increases. The ERS with the group-specific start already begins with a relatively smaller *MSE*, which means that the ability estimation is accurate even with a small number of items answered. However, the ERS with the cold start produces considerably higher *MSE*s than the group-specific approach in the first session. In each panel, the performances of the ERS with the two starting values coincide after 20 items at least. The results suggest that using the group-specific start is capable of enhancing ability estimation without having students answer an excessive number of items.

Study 2: When there is true ability change between sessions

Study design

In the second simulation study, we focused on exploring the effects of addressing the ability changes between sessions, assuming that the first part of study had been addressed. Therefore, the data generation model included four

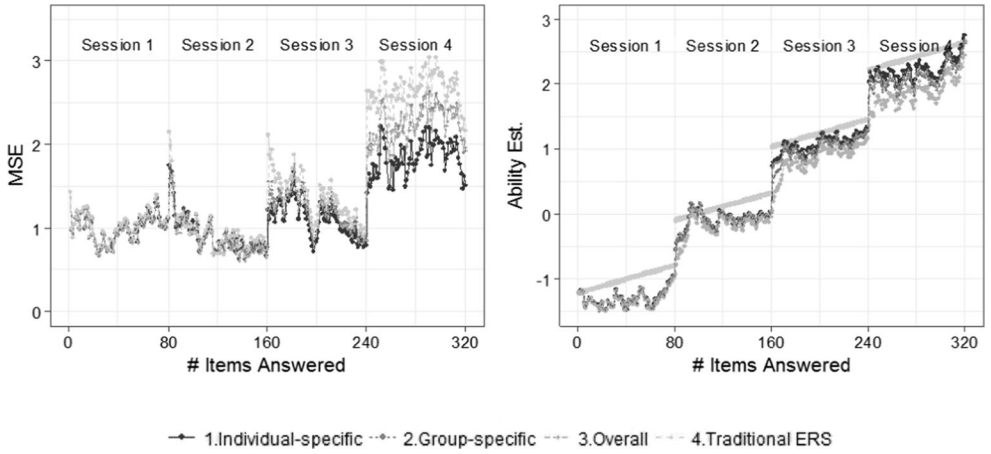

Fig. 5 *MSE*s (left) and estimated ability trajectories for an average student (right; straight lines indicate the true ability trajectory averaged for the group) when there is ability change with large student variation during the period between sessions. Note that with the validation data (n = 50 new students; 80 items per session), the initial ability estimates in the ERS for the students were determined by group-specific starting values

components: (a) initial ability, (b) the ability change while engaged in the learning environment, (c) the ability change while not engaged in the learning environment, and (d) a quadratic term for the ability change while not engaged in the learning environment. The quadratic term allowed us to simulate the possible acceleration of ability change over time while participants were not engaged in the learning environment. Therefore, a mathematical formulation for the data generation model is as follows:

$$\begin{aligned} logit\left(P\left[Y_{pi(t)}=1\right]\right) &= \left(\alpha_{01}Z_{p1}+\alpha_{02}Z_{p2}+w_{0p}\right)+\left(\alpha_{11}Z_{p1}+\alpha_{12}Z_{p2}+w_{1p}\right)wtime_{p(t)} \\ &+\left(\alpha_{21}Z_{p1}+\alpha_{22}Z_{p2}+w_{2p}\right)btime_{p(t)} \\ &+\left(\alpha_{31}Z_{p1}+\alpha_{32}Z_{p2}+w_{3p}\right)btime^{2}_{p(t)}-\beta_{p(t)}. \end{aligned} \quad (9)$$

Between-session change Four scenarios are considered in regard of fixed effects of the between-session parameters given α_{21} = .7 and α_{22} = .1 for nondyslexic and dyslexic groups. The four scenarios are varied by random effects and numbers of items as follows:

- Scenario A: with between-session ability change, small student variation (σ_{w0} = 0.5, σ_{w2} = 0.2), and 40 items per session
- Scenario B: with between-session ability change, large student variation (σ_{w0} = 1.5, σ_{w2} = 1.5), and 40 items per session
- Scenario C: with between-session ability change, small student variation (σ_{w0} = 0.5, σ_{w2} = 0.2), and 80 items per session
- Scenario D: with between-session ability change, large student variation (σ_{w0} = 1.5, σ_{w2} = 1.5), and 80 items per session

The between-session time (i.e., $btime_{pi}$) was generated from a Poisson distribution with a mean and variance of 1 (per day).

Within-session change In any of the data generation scenarios, the true fixed and random effects for the within-session period were determined as follows: α_{11} = .18, α_{12} = .05, and σ_{w2} = .0008. The within-session time (i.e., $wtime_{p(t)}$) was generated on the basis of approximately 0.05 h for solving each item in the session.

Table 3 Parameter estimates of the explanatory IRT models (n = 250 training set): Between-session ability change with small student variation

Parameter	True	Overall		Group-Specific	
		Est.	*SE*	Est.	*SE*
Intercept		– 0.810	0.100		
α_{01} (Nondyslexia)	0.5			0.390	0.112
α_{02} (Dyslexia)	– 2.5			– 2.217	0.178
Between-Session Slopes		0.446	0.090		
α_{21} (Nondyslexia)	0.7			0.681	0.133
α_{22} (Dyslexia)	0.1			0.245	0.110

Table 4 Parameter estimates of the explanatory IRT models (n = 250 training set): Between-session ability change with large student variation (40 items per session)

Parameter	True	Overall		Group-Specific	
		Est.	*SE*	Est.	*SE*
Intercept		– 0.760	0.084		
α_{01} (Nondyslexia)	0.5			0.364	0.108
α_{02} (Dyslexia)	– 2.5			– 1.912	0.125
Between-Session Slopes		0.428	0.089		
α_{21} (Nondyslexia)	0.7			0.567	0.126
α_{22} (Dyslexia)	0.1			0.284	0.123

Table 5 Parameter estimates of the explanatory IRT models (n = 250 training set): Between-session ability change with small student variation (80 items per session)

Parameter	True	Overall		Group-Specific	
		Est.	*SE*	Est.	*SE*
Intercept		– 1.001	0.107		
α_{01} (Nondyslexia)	0.5			0.448	0.105
α_{02} (Dyslexia)	– 2.5			– 2.137	0.116
Between-Session Slopes		0.405	0.084		
α_{21} (Nondyslexia)	0.7			0.650	0.113
α_{22} (Dyslexia)	0.1			0.238	0.126

Item difficulty As in Study 1, $\beta_{p(t)}$ was randomly drawn from $N(0, 2)$, and the item responses were randomly generated on the basis of a binomial distribution with the probability from the Eq. 9.

Starting values To determine starting values that take into account ability changes between sessions, four options were considered in this study. The first option (called "Traditional ERS" in Figs. 2, 3, 4, and 5 below) was to continue with the last ability estimate from the previous session. The second option ("Overall" in Figs. 2, 3, 4, and 5) was to use overall slopes that do not differentiate between the groups; therefore, the data-analyzing model did not include group indicators—that is, logit $P[Y_{pi(t)} = 1] = \left(\sum_{g=1}^{2}\alpha_{0g}Z_{pg} + w_{0p}\right) + \left(\sum_{g=1}^{2}\alpha_{1g}Z_{pg} + w_{1p}\right)wtime_{pt} + \left(\sum_{g=1}^{2}\alpha_{2g}Z_{pg} + w_{2p}\right)btime_{pt} - \beta_i$.

The third option ("Group-Specific" in Figs. 2, 3, 4, and 5) was to use the group-specific slopes corresponding to dyslexic or nondyslexic groups (i.e., α_{2g}, g = 1, 2). Finally the fourth option ("Individual-Specific" in Figs. 2, 3, 4, and 5) was to use each individual student's deviation (i.e., w_{2p}, p = 1, …, P) as well as the group-specific averages. The data-analyzing model was equivalent to that in Eq. 7. Note that the accuracy of the students' deviations improves as more data for the student are collected throughout longer study sessions.

Table 6 Parameter estimates of the explanatory IRT models (n = 250 training set): Between-session ability change with large student variation (80 items per session)

Parameter	True	Overall		Group-Specific	
		Est.	*SE*	Est.	*SE*
Intercept		– 0.824	0.100		
α_{01} (Nondyslexia)	0.5			0.311	0.105
α_{02} (Dyslexia)	– 2.5			– 2.310	0.128
Between-Session Slopes		0.433	0.083		
α_{21} (Nondyslexia)	0.7			0.674	0.107
α_{22} (Dyslexia)	0.1			0.193	0.118

Table 7 Parameter estimates of the IRT models for the online data (n = 500)

Parameter	Empty model		Explanatory model	
	Est.	*SE*	Est.	*SE*
Intercept	0.094	0.090	– 0.732	0.155
Type of learning			– 0.302	0.095
Grade			0.315	0.035
Gender			– 0.039	0.146
Between-Session Slopes	0.013	0.063	0.019	0.176
Type of learning			– 0.008	0.007
Grade			– 0.001	0.023
Gender			0.017	0.127

Results

The results are summarized for each of the four data generation scenarios.

Between-session ability change with small student variation (40 items per session) The first scenario assumed that the ability level of students generally evolves between sessions (α_{21} = 0.7, α_{21} = 0.1), with small variation in ability change among students (σ_{w2} = 0.2). It also assumed small variation among students in their initial ability levels (σ_{w0} = 0.5). Therefore, the scenario led to 86% of students who experienced constructive learning, and the remaining 14% of students who experienced forgetting between sessions. Table 3 presents the parameter estimates from the explanatory IRTs for the training data (n = 250 students). Table 3 shows that the overall slope estimate without using the grouping variable is 0.446. On the other hand, the group-specific estimates are approximately 0.681 and 0.245 for the nondyslexic and dyslexic groups, respectively.

With the validation data (n = 50 new students), the ERS with the choice of group-specific starting values as initial ability levels was implemented. Figure 2 visualizes *MSE* as a function of the total number of items answered at each iteration step in the ERS (left) and the true and estimated ability trajectories for an average student (right). The *MSE* value at the beginning is around 0.6 and decreases within Session 1. Such a decreasing trend in *MSE* can also be found within Sessions 2–4, meaning that ability estimation in the ERS gets more accurate as more items are given to students. However, it is noticeable that the *MSE* tends to take considerable leaps between consecutive sessions. This is because there are true systematic ability changes during these periods (α_{20} = 0.7, α_{21} = 0.1), but also unsystematic (individual-specific) ability changes. When methods using the explanatory IRT model

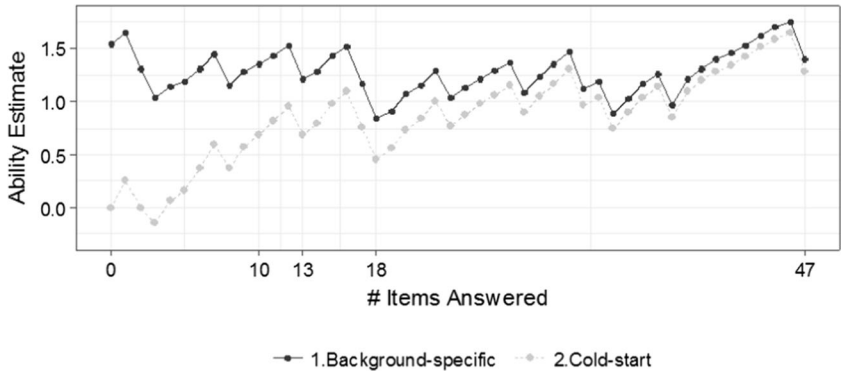

Fig. 6 Example of ability estimates for a student, given different start values in the ERS. Note that in this figure, the between-session effect is not addressed (see Fig. 7)

(in the "Individual-specific" and "Group-specific" forms) are used, the *MSE* values are smaller than in the traditional ERS—that is, because the explanatory IRT model succeeds in accounting for those changes.

Also, the right panel of the figure shows bias—that is, the gap between the true and estimated ability trajectories. The results suggest that methods using the explanatory IRT model (both "Individual-specific" and "Group-specific") consistently outperform the traditional ERS because IRT enables us to reduce bias due to the ability change between sessions.

Between-session ability change with large student variation (40 items per session) The second scenario assumed that the ability level of students generally evolves between sessions (α_{21} = 0.7, α_{21} = 0.1), but with large variation in ability change among students (σ_{w2} = 1.5). Also, we assumed large variation among students in their initial ability levels (σ_{w0} = 1.5). Therefore, the scenario led to 60% of students who experienced constructive learning, and the remaining 40% of students who experienced forgetting between sessions. Table 4 presents parameter estimates from the explanatory IRTs for the training data (n = 250 students). The results include posterior means and standard deviations from the models fit with the without using a background variable—that is, dyslexia. The table shows that the overall slope estimate without using the grouping variable is 0.428. On the other hand, the group-specific estimates are approximately 0.567 and 0.284 for nondyslexic and dyslexic groups, respectively.

Figure 3 visualizes *MSE* as a function of the total number of items answered at each iteration step in the ERS (left) and the true and estimated ability trajectories for an average student (right). As compared to Scenario A, the *MSE* values are greater, in general. The *MSE* value at the beginning is around 1.3, and it reduces almost to 0.4 within Session 1. Such a decreasing trend in *MSE* can also be found within Sessions 2–4, meaning that the ability estimation in ERS gets more accurate as more items are given to students. However, it is noticeable that the *MSE* tends to take considerable leaps between two consecutive sessions, due to the true systematic (α_{20} = 0.7, α_{21} = 0.1) and unsystematic (individual-specific) ability changes during these periods. Therefore, when methods using the explanatory IRT model (in the "Individual-specific" and "Group-specific" versions) are used, the *MSE* values are smaller than in the traditional ERS. Between the two approaches, the individual-specific estimate seems to perform better, especially in Session 4. Also, the right panel of the figure shows bias—that is, the gap between the true and estimated ability trajectories. The results suggest that methods using the explanatory IRT model (both "Individual-specific" and "Group-specific") consistently outperform the traditional ERS, because IRT enables us to account for the ability change between sessions more accurately.

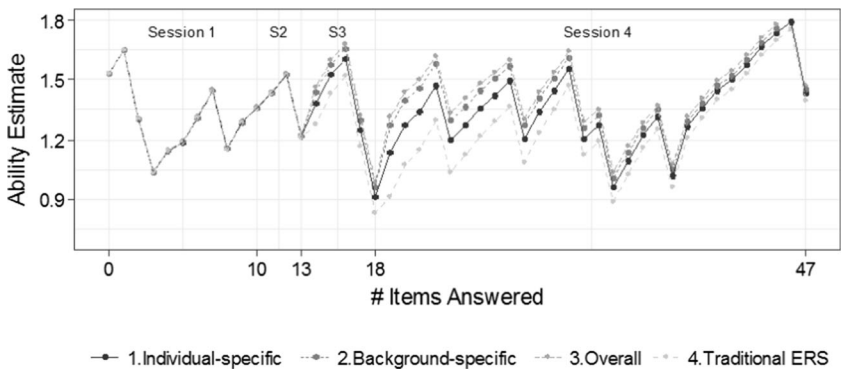

Fig. 7 Example of ability estimates for a student with and without addressing the between-session effect in the ERS

Between-session ability change with small student variation (80 items per session) As in Scenario A, we assumed that the ability levels of students generally tend to evolve between sessions ($\alpha_{20} = 0.7$, $\alpha_{21} = 0.1$) with small variation among students ($\sigma_{w2} = 0.2$). It also assumed that variation among students in their initial ability levels was small ($\sigma_{w0} = 0.5$). However, instead of 40 items per session, we simulated student responses to 80 items per session. Therefore, the scenario led to 82% of students who experienced constructive learning, and the remaining 18% of students who experienced forgetting between sessions. Table 5 presents parameter estimates from the explanatory IRTs for the training data ($n = 250$ students). The results include posterior means ("Est.") and standard deviations ("*SE*") from the models fit with or without using a background variable—that is, dyslexia. The table shows that the overall slope estimate without using the grouping variable is 0.405. On the other hand, the group-specific estimates are approximately 0.650 and 0.238 for the nondyslexic and dyslexic groups, respectively.

Figure 4 visualizes *MSE* as a function of the total number of items answered at each iteration step in the ERS (left) and the true and estimated ability trajectories for an average student (right). During Session 1, the *MSE* begins at 0.6 and moves to almost zero at the end of Session 1. During Sessions 3 and 4, noticeable gaps appear between individual- or group-specific approaches, as compared to the traditional ERS approach (in light gray). The two approaches reduce *MSE* in leaps between Sessions 1 and 2 (Items 80 and 81), Sessions 2 and 3 (Items 160 and 161), and Sessions 3 and 4 (Items 240 and 241). In this scenario, the individual-specific approach works best but does not particularly outperform the group-specific estimate. Also, the right panel of the figure shows bias—that is, the gap between the true and estimated ability trajectories. The results suggest that the methods using the explanatory IRT model (both "Individual-specific" and "Group-specific") consistently outperform the traditional ERS, because IRT enables us to account for ability change between sessions more accurately.

Between-session ability change with large student variation (80 items per session) In the last scenario, we considered a case similar to Scenario B, in which the ability levels of students generally evolve between sessions ($\alpha_{20} = 0.7$, $\alpha_{21} = 0.1$), with large variation among students ($\sigma_{w2} = 1.5$). It also assumes small variation among students in their initial ability levels ($\sigma_{w0} = 1.5$). However, instead of 40 items per session, we simulated student responses to 80 items per session. Therefore, the scenario led to 74% of students who experienced constructive learning, and the remaining 26% of students who experienced forgetting between sessions. Table 6 shows that the overall slope estimate without using the grouping variable is 0.433. On the other hand, the group-specific estimates are approximately 0.674 and 0.193 for nondyslexic and dyslexic groups, respectively.

Figure 5 visualizes *MSE* as a function of the total number of items answered at each iteration step in the ERS (left) and the true and estimated ability trajectories for an average student (right). Due to the large variation in ability change between sessions, we found considerable leaps between Sessions 1 and 2 (Items 80 and 81), Sessions 2 and 3 (Items 160 and 161), and Sessions 3 and 4 (Items 240 and 241). Although all four approaches result in very similar performance during Session 2, the individual-specific and group-specific approaches outperform during Sessions 3 and 4. In particular, the individual-specific approach produces smaller *MSE*s than the group-specific approach, and the gap becomes bigger in Session 4. Also, the right panel of the figure shows bias—that is, the gap between the true and estimated ability trajectories. The results suggest that the methods using the explanatory IRT model (both "Individual-specific" and "Group-specific") consistently outperform the traditional ERS, because IRT enables us to account for the ability change between sessions more accurately.

Real-data example

Data description

For illustrative purposes, we used a data set collected from a Web-based learning platform, "Oefenweb" (Oefenweb.nl). It was designed as an item-based e-learning environment for 200,000 pupils from a total of 2,000 mainly primary schools in the Netherlands. The learning environment supplements children's cognitive development in math and language at their own ability level, and teachers can receive information in order to get informed about their students. We used data obtained from one of its exercise programs, called Math Garden. In particular, the present data refer to math exercises related to the addition operation, which were collected between Fall 2016 and Spring 2017, during one school year. Those exercises were developed on the basis of the Rasch model framework. The platform includes data from large numbers of students and learning items—specifically, 1,562 students' responses to items across four study sessions in total. Note that the total number of items for the four study sessions varied by students—the average number of items per student was 81.3, with a minimum of 8 and a maximum of 576—and therefore, the data set includes missing responses. In addition to the student responses to the items, the data set also contains background information variables for the students, including (a) type of learning (0 = *easy*, 1 = *moderate*, and 2 = *challenging*), (b) grade (3rd, 4th, 5th, and 6th graders), and (c) gender (0 = *female*, 1 = *male*). Finally, the data set also contains timestamps recording when the students start and finish solving each item, so the amounts of time spent within (*wtime*) and outside (*btime*) the study sessions can be

calculated. In particular, *wtime* indicates measurement time points in hours across all study sessions, and *btime* shows the spacing time in days between study sessions, computed from the session-specific time stamps. The median spacing time between two consecutive study sessions was 1.93 days. Among the students, 500 were selected for analyzing with explanatory IRT models, and the remaining 1,062 students were used to keep track of ability trajectory using the traditional ERS. On the basis of this setup, the same procedures as described in Studies 1 and 2 were applied.

Results

Table 7 demonstrates the estimated model parameters from fitting explanatory IRT models to the Math Garden data. The middle columns of the table (called "Empty Model") show the estimated parameters and standard errors (posterior means and standard deviations) from fitting an IRT model that included the overall intercept and the between-session slope (for *btime* variable). At this stage, no student characteristics were included as explanatory variables. The right columns of Table 7 (called "Explanatory Model") show estimated parameters and standard errors with respect to the intercept and the between-session slope as moderated by a set of explanatory variables that characterize the students. Similar to the simulation study (see Study 1), the estimated intercept(s) from these two models produce the starting values for the ERS to begin tracing ability growth, particularly for students who have newly engaged with the environment.

Figures 6 and 7 show estimated ability trajectories for a randomly chosen student. As in the simulation results (i.e., Study 1), Fig. 6 presents the impact of choosing starting values when the student came in to the learning environment for the first time. In this figure, the between-session effect is not addressed, which will be depicted in Fig. 7. As in Study 1, two options were considered: starting values estimated using all explanatory variables (see the right columns of Table 7) or a cold start, assuming that all learners' ability levels are equivalently zero. The figure shows that the starting values estimated using students' background are greater than the cold start by approximately 1.5 points. It is noticeable that the gap between the two options gets narrower as the student answers more items. In particular, the gap becomes negligible after 32 items answered. That implies that ignoring students' characteristics may cause considerable bias in ability trajectory estimation when new users come in, and therefore suboptimal adaptivity of the learning environment.

For the same student, Fig. 7 compares the estimated ability trajectories with and without considering ability change during the period between sessions (i.e., while not engaged with the learning environment) for the ERS. To avoid the initial cold-start problem, the background-specific estimates from Fig. 6 were used. Because the four options only differ in the ways the between-session effect is accounted for, the curves coincide in Session 1. Also, the differences are extremely minor for Session 2 ("S2"). At the beginning of Session 3 ("S3"), the ability trajectory using the three model-based starting values tends to be greater than with the traditional ERS. Finally, in Session 4, the ability estimates when using the three model-based starting values again tend to be greater than with the traditional one. In particular, the overall starting value gave the highest value, followed by the background-specific starting value and the individual-specific starting value. The estimates by using individual-specific starting values are located between the other two model-based starting values and the traditional ERS. After a longer sequence of items within the session, however, the gaps among the four approaches tend to be negligible.

Discussion and conclusion

The present study proposed methods to address the cold-start problem in e-learning environments by implementing the explanatory IRT model with the ERS. The proposed methods were empirically evaluated via a simulation study. We considered various scenarios that differed from each other in (a) whether the group-specific initial ability estimates were specified for new students or (b) whether group-specific and/or individual-specific effects of the explanatory variables were included for the students' ability change while not using the learning environment. Those explanatory IRT models were evaluated under conditions in which the true initial ability parameters were different across multiple groups, the true ability trajectory between sessions was either positive or negative, and relatively small or large variances across students were generated. The proposed models were evaluated with the cross-validation technique, in which the models were built using part of the data set (a training set) and tested on the remaining part of the data set (a testing set, which we considered to be data from new students). Finally, the proposed methods were illustrated using real data obtained from an e-learning environment.

The results of the study showed that the ERS with the explanatory IRT models provided better latent ability estimates when a student entered the e-learning environment than did the traditional ERS with the Rasch model. The *MSE* values for the ability estimates at the initial stage were consistently lower for the explanatory models, and similar findings were observed across measurement occasions. These findings imply that in order to obtain more accurate ability estimates in the ERS, an explanatory IRT model that can take students' information into account should be implemented. The flexibility to include the explanatory variables plays a significant role in the ERS, and individualized initial estimates could increase the efficiency of the e-learning system. On the basis of these

findings, we recommend using explanatory IRT models with the ERS to obtain more accurate and efficient ability estimates.

In addition, on the basis of the present simulation study results, we found that when students resumed the e-learning system, group-specific abilities were the better estimates for the initial ability level at which to initiate the ERS, rather than the estimates at the last measurement occasion of the previous session. In the present study, explanatory IRT models that included between-session variables showed more accurate ability estimates than did the model without the between-session variables when there was a between-session effect. The model with the between-session variables also outperformed for conditions in which the between-session effects were both positive and negative. These findings imply that for e-learning environments, the session-specific approach should be encouraged to estimate the between-session effect.

It is also worthwhile to note that when the explanatory IRT models were estimated, we employed the Bayesian estimation method. In general, for situations in which the fitted model becomes more complex or data sets include large number of observations, it may not be uncommon to observe nonconvergence of the parameter estimation, or it may take a much longer runtime for the model to converge. A similar issue occurred in a study in which longitudinal growth IRT modeling for the e-learning data was examined (Kadengye, Ceulemans, & Van den Noortgate, 2015). Given that the Bayesian approach can be an alternative to obtaining estimates by sampling from the posterior distribution, we initially estimated the models using both a numerical integration approach and MCMC estimations, and found that MCMC yielded more accurate estimates and took less runtime. Although more studies will be needed to precisely compare both estimation methods for explanatory IRT models in the context of the e-learning environment, our study suggests that the Bayesian approach toward explanatory IRT models works well.

We also acknowledge some limitations of the study. For the simulation study, we only incorporated one categorical explanatory variable to address the cold-start problem. However, in practical settings more student information will be available and collected, as is shown in the real-data illustration, and those background variables should be included in the explanatory IRT model. Including more variables is expected to lead to an increase in precision, but it also increases the complexity of the explanatory IRT model, and the same pattern as in the current approaches may not be obtained. In addition, the ERS that we considered in this study does not assume adaptive item sequencing. In other words, in the present simulation study, the items were generated randomly across measurement occasions for each student, and the item difficulty parameters were assumed to be fixed and considered known (i.e., estimated using a large prior calibration study). However, given that the ERS allows for updating not only the ability estimate but also the item difficulties simultaneously, it would be worthwhile to investigate the performance of the explanatory IRT model with more complex conditions (e.g., when item difficulties cannot be considered known, and therefore when the cold-start problem also applies to the item side). Finally, the present work demonstrates the added value of using the explanatory IRT method on top of the traditional method (Glickman, 1999) of addressing the cold-start problem by using large step sizes at the beginning of new sessions. Specifically, we took a bigger step size to start the ERS algorithm (by setting $K = 0.7$ at the beginning of each study session),and linearly decreased the size as a function of the total number of items answered. More research exploring the optimal step size function will be desirable. For example, Klinkenberg et al. (2011, p. 1816) followed Glickman's reasoning and proposed making K a function of the number of items answered and the elapsed time between two answers.

Nonetheless, the results of this study provide valuable information about how to deal with the cold-start problem in e-learning environments. We recommend fitting explanatory IRT models that can be used to get good initial starting values for new students, or to get improved starting values at the start of new sessions. These (relatively computing-intensive) analyses do not have to be done on the fly, but can be repeated to get more precise estimates of model parameters when the amount of available data increases. Also, the analyses to estimate student-specific trajectories do not have to be done on the fly. Ideally these should be performed between study sessions. One could, for instance, update the estimates once a day. The study is expected to allay concerns about the ERS and catalyze the usefulness of the e-learning system in educational settings. Furthermore, we acknowledge that the cold start is a problem that can be encountered in any educational setting—for example, when there is a new student (or new learning materials are introduced) in a traditional classroom setting. It is likely that the teacher will try to overcome this cold-start problem using the same principle that underlies our explanatory IRT approach: by predicting ability using the background characteristics of the student. The advantage of the classroom setting is that the teacher in principle could use multiple types of data (e.g., the attitude and participation of the student) to estimate the extent to which the student has mastered the content. In contrast, a disadvantage is that the predictions are not necessarily based on an objective, evidence-based model, but rather may be prone to prejudices. In addition, teachers cannot observe all students permanently, so they may miss important information that could be used to update their ability estimates. We hope this study will inspire researchers in face-to-face educational settings, too.

Acknowledgements This research include a methodological approach and a real data example from the LEarning analytics for AdaPtive Support (LEAPS) project, funded by imec (Kapeldreef 75, B-3001,

Leuven, Belgium) and the Agentschap Innoveren & Ondernemen. The LEAPS project aimed to develop a self-learning analytical system to enable adaptive learning. This support system can be integrated into educational games and in supporting software for difficult readers and for professional communication. Partners from a broad field of expertise work together within the consortium. Examples include educational and cognitive scientists, software developers, statisticians, experts in human-computer interaction and educational publishers. This extensive interdisciplinary collaboration makes it possible to create a much-needed and commercial solution.

References

Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). Generalization of recommender systems: Collaborative filtering extended to groups of users and restricted to groups of items. *Expert Systems with Applications*, *39*, 172–186.

Brinkhuis, M. J., & Maris, G. (2009). *Dynamic parameter estimation in student monitoring systems* (Measurement and Research Department Reports, Rep. No. 2009-1). Arnhem, The Netherlands: Cito.

Brusilovsky, P., & Peylo, C. (2003). Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education*, *13*, 159–172.

Coomans, F., Hofman, A., Brinkhuis, M., van der Maas, H. L., & Maris, G. (2016). Distinguishing fast and slow processes in accuracy-response time data. *PLoS ONE*, *11*, e0155149. https://doi.org/10.1371/journal.pone.0155149

Dai, Y., & Mislevy, R. (2009). *A mixture Rasch model with a covariate: A simulation study via Bayesian Markov chain Monte Carlo estimation* (PhD dissertation). College Park, MD: University of Maryland. Retrieved from http://drum.lib.umd.edu/handle/1903/9926

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models*. New York, NY: Springer.

Elo, A. E. (1978). *The rating of chessplayers, past and present* (Vol. 3). London, UK: Batsford.

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, *47*, 432–457.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, *25*, 165–173. https://doi.org/10.2307/271064

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, *48*, 377–394.

Humphrey, N., & Mullins, P. M. (2002). Research section: Personal constructs and attribution for academic success and failure in dyslexia. *British Journal of Special Education*, *29*, 196–203.

Kadengye, D. T., Ceulemans, E., & Van den Noortgate, W. (2014). A generalized longitudinal mixture IRT model for measuring differential growth in learning environments. *Behavior Research Methods*, *46*, 823–840. https://doi.org/10.3758/s13428-013-0413-3

Kadengye, D. T., Ceulemans, E., & Van den Noortgate, W. (2015). Modeling growth in electronic learning environments using a longitudinal random item response model. *Journal of Experimental Education*, *83*, 175–202.

Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology Research and Development*, *53*, 83–93.

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. https://doi.org/10.1016/j.compedu.2011.02.003

Maris, G., & van der Maas, H. (2012). Speed–accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633.

Papousek, J., Pelánek, R., & Stanislav, V. (2014, July). *Adaptive practice of facts in domains with varied prior knowledge*. Paper presented at the Educational Data Mining 2014 Conference, London, UK.

Pereira, A. L. V., & Hruschka, E. R. (2015). Simultaneous co-clustering and learning to address the cold start problem in recommender systems. *Knowledge-Based Systems*, *82*, 11–19.

Plummer, M. (2015) Just another Gibbs sampler (JAGS) [Software]. Retrieved from http://mcmc-jags.sourceforge.net

Polychroni, F., Koukoura, K., & Anagnostou, I. (2006). Academic self-concept, reading attitudes and approaches to learning of children with dyslexia: Do they differ from their peers? *European Journal of Special Needs Education*, *21*, 415–430.

R Core Team. (2013). R: A language and environment for statistical computing (Version 3.3.3). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Savi, A. O., van der Maas, H. L., & Maris, G. K. (2015). Navigating massive open online courses. *Science*, *347*, 958–958.

Shute, V., & Towle, B. (2003). Adaptive e-learning. *Educational Psychologist*, *38*, 105–114.

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher*, *18*, 8–14.

Snow, R. E. (1996). Aptitude development and education. *Psychology, Public Policy, and Law*, *2*, 536–560. https://doi.org/10.1037/1076-8971.2.3-4.536

Su, Y. S., & Yajima, M. (2015). R2jags: Using R to run "JAGS" (R package version 0.5-7). Retrieved from

Tang, T., & McCalla, G. (2004, August). *Utilizing artificial learners to help overcome the cold-start problem in a pedagogically-oriented paper recommendation system*. Paper presented at the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Eindhoven, The Netherlands.

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.

Vukovic, R. K., Lesaux, N. K., & Siegel, L. S. (2010). The mathematics skills of children with reading difficulties. *Learning and Individual Differences*, *20*, 639–643.

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.

Wauters, K., Desmet, P., & Van den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, *58*, 1183–1193.