Measuring growth in students' proficiency in MOOCs: Two component dynamic extensions for the Rasch model

Dmitry Abbakumov^{1,2,3} · Piet Desmet^{1,4} · Wim Van den Noortgate^{1,2}

Published online: 27 September 2018
© Psychonomic Society, Inc. 2018

Abstract

Massive open online courses (MOOCs) are increasingly popular among students of various ages and at universities around the world. The main aim of a MOOC is growth in students' proficiency. That is why students, professors, and universities are interested in the accurate measurement of growth. Traditional psychometric approaches based on item response theory (IRT) assume that a student's proficiency is constant over time, and therefore are not well suited for measuring growth. In this study we sought to go beyond this assumption, by (a) proposing to measure two components of growth in proficiency in MOOCs; (b) applying this idea in two dynamic extensions of the most common IRT model, the Rasch model; (c) illustrating these extensions through analyses of logged data from three MOOCs; and (d) checking the quality of the extensions using a cross-validation procedure. We found that proficiency grows both across whole courses and within learning objectives. In addition, our dynamic extensions fit the data better than does the original Rasch model, and both extensions performed well, with an average accuracy of .763 in predicting students' responses from real MOOCs.

Keywords Psychometrics · Item response theory · Cross-classification multilevel logistic model · Learning effects

Massive open online courses (MOOCs), which emerged a decade ago, are a progressive phenomenon in distance education. A MOOC is a free online course available for anyone to enroll. For students of various ages, MOOCs are a flexible way to obtain new skills and to advance careers. For universities, MOOCs are an efficient way to deliver education at scale. Typically, a MOOC consists of prerecorded video lectures, reading assignments, assessments, and forums. There are several provider platforms on which universities publish MOOCs, including Coursera, edX, XuetangX, FutureLearn, Udacity, and MiriadaX. In December 2016, the estimated total number of MOOCs was 6,850, from over 700 universities around the world. Coursera, edX, and XuetangX are the largest MOOC provider platforms, with over 39 million learners (Shah, 2016).

The main aim of MOOCs, like any other learning environment, is a growth in students' proficiency. Growth tracking is essential for all parties involved: for students, to understand his or her progress in the proficiency level, and for professors, to infer how efficient a course is and/or to decide when to support a student (or advance him/her) through the course. Unfortunately, proficiency is a latent variable that cannot be observed directly; it can be estimated, however, on the basis of observable variables—for example, a student's performance on assessment items. Since proficiency itself is a latent variable, its growth is also a latent variable. To link the observable side to the latent side, there are specific rules, called *psychometric theories*.

Psychometric approaches based on item response theory (IRT) traditionally consider proficiency as a constant variable (Lord & Novick, 1968). These approaches were developed for measuring proficiency but not for measuring its change. For example, the most common psychometric model, the Rasch model (Lord & Novick, 1968; Rasch, 1960), states that

$$\text{Logit}(\pi_{ij}|\theta_j) = \ln(\pi_{ij}/1-\pi_{ij}) = \theta_j - \delta_i \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}). \quad (1)$$

In this model, Y_{ij} is the observable score of student j to an item i , which equals 1 for a correct response or 0 for an incorrect response. This variable therefore can be considered as being Bernoulli-distributed, with probability π_{ij} , which in turn is described by a logistic function of the difference between a static

✉ Dmitry Abbakumov
dmitry.abbakumov@kuleuven.be

¹ ITEC imec, Leuven, Belgium

² Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

³ eLearning Office, National Research University Higher School of Economics, Moscow, Russia

⁴ Faculty of Arts, KU Leuven, Leuven, Belgium

student's parameter (θ_j) and an item parameter (δ_i), which are often interpreted as proficiency and difficulty, respectively. The values for the proficiency parameter are typically considered as a random sample from a normal population distribution, with $\theta_j \sim N(0, \sigma_\theta^2)$, whereas the items are considered as a fixed set. Hence, the parameters that are estimated are the difficulty parameters and the variance between students in the proficiency parameter (the mean of the distribution of proficiencies is constrained to be zero, in order to make the model identified). The main assumptions of the Rasch model are *unidimensionality*, which means that only one kind of proficiency is measured by a set of items in a test, and *local independence*, which means that when the proficiency influencing test performance is held constant, students' responses to any pair of items are statistically independent (Hambleton, Swaminathan, & Rogers, 1991; Molenaar, 1995).

One of the common approaches to detect growth in a course is to administer the same test to a student, first at the beginning and then at the end of that course. Following classical test theory (CTT), the first score can be subtracted from the second score, and the resulting difference is used as a value of the student's growth (Davis, 1964). According to the IRT tradition, however, the value of growth comes from the difference in the proficiency estimates obtained on two measurement occasions (Andersen, 1985). IRT makes it possible to use sets of items from two measurement occasions that only partially overlap and in general are not equally difficult. Such overlapping allows for placing items from both sets on a common scale (Hambleton et al., 1991), and as a result, getting a difference in proficiency estimates between the beginning and end of the course. When we can make use of items that have previously been calibrated, we can even provide items to new students at the two measurement occasions that do not overlap at all.

Yet, psychometricians have successfully attempted to model dynamic processes within the framework of the Rasch model. The first class of such models focuses on assessments, the second on learning environments. The first class might be decomposed further in two subclasses that differ from each other in whether the change in students' proficiency between or within assessments is modeled.

Fisher (1976, 1995) presented linear logistic models that measure the change in proficiency between assessments—for instance, between a pretest and a posttest. These models are based on the idea that an item given to the same student at two different time points can be considered a pair of items, with two different item difficulty parameters. Thus, any change in proficiency occurring between the measurement occasions is described through a change of the item parameters. During the assessment, proficiency is assumed to remain constant.

Another subclass of models aims at measuring the change in proficiency within the assessment—for instance, due to a

learning effect. Verhelst and Glas (1993, 1995) proposed a dynamic generalization for the Rasch model:

$$\text{Logit}(\pi_{ij}|\theta_j) = \ln(\pi_{ij}/1-\pi_{ij}) = \theta_j + t_{ji}\gamma - \delta_i \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}), \quad (2)$$

where t_{ji} is the number of correct answers or the number of items viewed by student j up to item $i - 1$. γ links t_{ji} with the probability of the correct answer, and therefore represents the growth. The presentation of t_{ji} depends on the dynamic process within the assessment that a researcher aims to estimate or control. For instance, if the only feedback displayed after a student's response is that the response was correct, it is assumed that students learn from correctly answered items only (Verguts & De Boeck, 2000). In this case, to control for proficiency growth within the assessment, a researcher can define t_{ji} as the number of correct answers. On the other hand, if more extended feedback is displayed after a student's response (e.g., hints or even the correct response), students may also learn from incorrect answers. Here, t_{ji} might be presented as the number of items viewed or even be decomposed into two variables, the number of correct answers and the number of incorrect answers, to control separately for growth from correctly answered items and from correct answers displayed after incorrect responses. Despite its flexibility, the model is limited, as it does not represent individual differences in growth; it indicates only the average learning trend. To overcome this limitation, it has been proposed to allow the growth effect γ to vary from student to student (De Boeck et al., 2011; Verguts & De Boeck, 2000). The resulting model is

$$\text{Logit}(\pi_{ij}|\theta_j) = \ln(\pi_{ij}/1-\pi_{ij}) = \theta_j + t_{ji}\gamma_j - \delta_i \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}), \quad (3)$$

where the students' growth parameters are considered as a random sample from a population in which the growth parameters are normally distributed, $\gamma_j \sim N(0, \sigma_\gamma^2)$.

The second class of models focuses on learning environments. Kadengye, Ceulemans, and Van den Noortgate (2014, 2015) proposed longitudinal IRT models, in which proficiency is considered as a function of both time within learning sessions and time between learning sessions—for instance,

$$\text{Logit}(\pi_{ij}) = (\alpha_0 + \omega_{0j}) + (\alpha_1 + \omega_{1j}) * wtime_{ij} + (\alpha_2 + \omega_{2j}) * btime_{ij} + v_i \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}), \quad (4)$$

where α_0 is the overall initial proficiency; ω_{0j} is the deviation of student j from α_0 ; $wtime_{ij}$ and $btime_{ij}$ are the amounts of time that passed while student j was, respectively, using and not using the learning environment, up to the moment of student j 's response to item i ; α_1 and α_2 are the overall population linear time trends within and between sessions, respectively; ω_{1j} and ω_{2j} are the deviations of student j from α_1 and α_2 , respectively, where student-specific random effects are

assumed to have a multivariate normal distribution; and v_i is a random item effect, with $v_i \sim N(0, \sigma_v^2)$. Thus, the authors introduce a dynamic concept $\theta_{ij} = (\alpha_0 + \omega_{0j}) + (\alpha_1 + \omega_{1j}) * wtime_{ij} + (\alpha_2 + \omega_{2j}) * btime_{ij}$, which corresponds to the proficiency of student j at the moment of responding to item i .

The direct use of the approaches presented above for measuring the growth in students' proficiency in MOOCs is hampered by a few challenges. Assessments have no common items, which makes the use of a scaling approach impossible. In addition, the time that a student spends in a MOOC is often not logged, which complicates the use of time-based approaches. Also the MOOC instructional design impedes the use of the models mentioned above. First, video lectures are considered the central instructional tool of a MOOC to support the learning objectives and prepare students for the associated assessments (Coursera, n.d.). Second, the way that a student interacts with an assessment item in MOOCs is complex: He/she can make several attempts to solve an item, and after a wrong response, he/she receives tips aimed at facilitating learning. This means that proficiency is expected to change mainly from the video lectures, but also from interaction with the items in assessments. We believe that the psychometric approach to modeling growth in MOOCs could be improved by taking account of information from these complex student–content interactions.

This study extends the growing research domain of modeling dynamic processes with IRT by focusing on the specificity and data structure of MOOCs. We propose to incorporate two novel growth trends, from video lectures and from interaction with the items in assessments, into the IRT framework in order to estimate differential latent growth that might be present in MOOC datasets. We realize this idea through two dynamic extensions for the Rasch model. Next, we illustrate these extensions using the data from three MOOCs. Finally, we check the performance of these extensions using a cross-validation procedure applied to data from these MOOCs. We expect that these extensions will provide possibilities for a more accurate estimate of students' latent traits in MOOCs.

Measuring two components of growth in proficiency in MOOCs

In this study, we start with the structure of MOOCs, specifically for Coursera courses. These are composed of modules, each of which is structured around a cohesive subtopic and typically lasts for one week. Each module consists of a set of lessons. Typically a lesson is structured around one or two learning objectives and includes several video lectures, which might be accompanied by additional instructional content—for instance, reading material, forum discussions, and practice

items. Each video lecture lasts 4–9 min. It takes a student about 30 min to complete a lesson. Each module is concluded by a summative assessment, which is realized as a 10- to 15-item test, a programming task, or a peer-review task. In this study we focus on the tests, which are the most popular type of assessment. The items in tests are typically multiple-choice or open-ended questions in which a student is expected to choose an option or options or respond with a number, sequence, word, or phrase.

Growth through the course During the course, a student watches video lectures in order to master learning objectives assigned for a certain lesson in a certain module. Typically a student has freedom in interacting with the video lectures. He or she can watch or skip a certain video lecture, which means that each student has an individual pattern of interaction with the video content, and the number of watched video lectures varies among students. To catch the growth in students' proficiency from video lectures, we can place all video lectures and all summative assessments of the course in one line successively—for instance, a sequence of video lectures in the first week, the summative assessment items from the first week, a sequence of video lectures in the second week, and so on. Now we can count a progressive sum of the video lectures (the observable variable) that a student watched before a certain summative assessment. In this case, the effect of the progressive sum on the students' performance with the summative assessment items represents the continuous growth in the student's proficiency (the latent variable) from the video content. Accumulation from the watched video lectures will probably boost a student's chances of making correct response on a certain summative assessment item. We remember that according to the MOOC design, video lectures are the core instructional tool. Thus, the growth in students' proficiency from the video lectures alone might be considered as the growth through the course.

Growth within a certain learning objective As we mentioned above, in MOOCs students can make several attempts to answer each item in an assessment. Thus, after a wrong response, the student may analyze his/her mistake, use the hint if one is assigned for the item, review the video lecture and notes, consult on forums, or even use extra-curricular materials, and afterward make a second attempt. In this case, the added activity will probably lead to an increased chance of making the correct response on that item from one attempt to the next, and we can catch local growth within a certain learning objective or even its part. However, it is important to note that the increasing chances might be explained by specific strategies of interacting with an item that may also be chosen by a student—clicking repeatedly on alternative

options in multiple-choice items, or guessing in open-ended items. In this case, we have no real growth (or pseudo-growth) within a certain learning objective.

In the following section, we model, visualize, and explain these dynamic concepts.

Two-component dynamic extensions for the Rasch model

As the basis of the two-component dynamic extensions, we use a reformulation of the Rasch model presented by Van den Noortgate, De Boeck, and Meulders (2003):

$$\text{Logit}(\pi_{ij}) = b_0 + u_{1j} + u_{2i} \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}), \quad (5)$$

where $u_{1j} \sim N(0, \sigma_{u1}^2)$ and $u_{2i} \sim N(0, \sigma_{u2}^2)$.

In this reformulation based on the principle of cross-classification multilevel models, which are generalizability theory models with logit-link functions, we have the intercept and two residual terms, referring to the student and the item, respectively. The means of both residual terms equal zero. Thus, the intercept equals the estimated logit of the probability of the correct response of an average student on an average item. The first residual term shows the deviation of the expected logit for student j from the overall logit. The higher this deviation, the higher is the probability of a correct response. Therefore, this residual term can be interpreted as the proficiency of student j , which equals θ_j from Eq. 1. The second residual term shows the deviation of the expected logit for item i from the mean logit, in the sense that the larger the residual, the higher the expected performance. The difficulty parameter δ_i from Eq. 1 is equivalent to $-(b_0 + u_{2i})$ from Eq. 4. Hence, the residual term u_{2i} refers to the relative easiness of item i , as compared to the mean easiness of all items, b_0 . The strength of this reformulation is that, in comparison to the original formulation of the Rasch model, the items are considered random variables, which makes the model very flexible for making extensions, because degrees of freedom are left that might include various item predictors (Van den Noortgate et al., 2003).

Extension with fixed growth effects

In the first extension, we transformed the growth trend proposed by Verhelst and Glas (1993, 1995) into the two components introduced above: the progressive sum of watched video lectures is used to model continuous growth through the course, and the response pattern for particular items in summative assessments is used to

model the local growth within a certain learning objective. Thus, in our first extension,

$$\text{Logit}(\pi_{ij}) = b_0 + b_1 * \text{video}_{ij} + b_2 * \text{attempt}_{ij} \quad (6)$$

$$+ u_{1j} + u_{2i} \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}),$$

b_0 equals the estimated logit of the probability of a correct response for an average student on an average item in the course summative assessments; video_{ij} is the progressive sum of the video lectures that student j watched before responding to item i , divided by 100 for better scaling; b_1 is the effect of the progressive sum, and is interpreted as the growth effect through the course; attempt_{ij} takes on values of 0, 1, 2, 3, or 4, for student j 's first, second, third, fourth, or fifth or higher attempt on item i , respectively; b_2 is the effect of attempt, and can be interpreted as the growth effect within a certain learning objective; and $u_{1j} \sim N(0, \sigma_{u1}^2)$ and $u_{2i} \sim N(0, \sigma_{u2}^2)$.

From Eq. 6 it can be deduced that the value $\theta_{0j} = u_{1j}$ corresponds to the initial proficiency of student j at the start of the course. In Fig. 1, this value is presented by the point with index 1. The value $\theta_{ij} = u_{1j} + b_1 * \text{video}_{ij}$ corresponds to the proficiency of student j at the moment of responding on item i . This dynamic value represents the continuous evolution of proficiency of student j through the whole course and determines his/her chances of a correct response on item i in the first, initial attempt. This value is presented on Fig. 1 by the point with index 2. However, if student j fails on this attempt, he/she would probably go on to a second attempt. The logit of the probability of a correct response in that case would increase by b_2 , which represents the local rise of the student's proficiency within a learning objective. This value is presented in Fig. 1 by the vertical segment with index 3.

Fig. 1 The point with index 1 presents the initial proficiency of a student. The point with index 2 corresponds to the proficiency of a student at the moment of responding on item i . The continuous growth of proficiency through the course is illustrated by the inclined dashed line. The local growth of proficiency within a certain learning objective is illustrated as the vertical segment with index 3

As in the model by Verhelst and Glas (1993, 1995), all students are assumed to have the same dynamics in their proficiency. We expect both growth effects, b_1 and b_2 , to be positive. This means that with each new video lecture watched, indicated by the progressive sum, and each new attempt made the chances of a correct response grow. We can thus explain growth by learning both throughout the course and within a certain learning objective.

Extension with random growth effects

The next extension follows the logic of modeling individual dynamics in proficiency that was presented in works by Verguts and De Boeck (2000), De Boeck et al. (2011), and Kadengye, Ceulemans, and Van den Noortgate (2014, 2015), by including the student-specific random effects. We assume that both effects introduced in the first extension can vary randomly from subject to subject. Thus, in our second extension,

$$\text{Logit}(\pi_{ij}) = b_0 + (b_{10} + b_{1j}) * \text{video}_{ij} + (b_{20} + b_{2j}) * \text{attempt}_{ij} \quad (7)$$

$$+ u_{1j} + u_{2j} \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}),$$

b_{10} is the overall effect of the progressive sum, the overall growth effect throughout the course, with b_{1j} as the deviation of the progressive sum effect for student j from the overall effect, thus defining the individual growth effect throughout a course; and b_{20} is the overall effect of attempt, or the overall growth effect within a certain learning objective, with b_{2j} as the deviation of the attempt effect for student j from the overall attempt effect, thus defining the individual growth effect within a certain learning objective. In the first version of this extension, u_{1j} , b_{1j} , and b_{2j} follow univariate normal distributions, $N(0, \sigma_{u1}^2)$, $N(0, \sigma_{b1}^2)$, and $N(0, \sigma_{b2}^2)$, respectively, whereas in the second version, u_{1j} , b_{1j} , and b_{2j} follow a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, with Σ as the variance–covariance matrix.

From Eq. 7, we can derive that the value $\theta_{0j} = u_{1j}$ represents the initial proficiency of student j at the start of the course (the point with index 1 in Fig. 1). The value $\theta_{ij} = u_{1j} + (b_{10} + b_{1j}) * \text{video}_{ij}$ corresponds to the proficiency of student j at the moment of responding on item i , which represents the continuous evolution of proficiency of student j through the whole course and determines his/her chances of a correct response on item i in the first, initial attempt (the point with index 2 in Fig. 1). In the case that student j fails in this attempt and goes on to a second, the logit of the probability of correct response would increase to $(b_{20} + b_{2j})$, which represents the local rise of the student's proficiency within a learning objective (the vertical segment with index 3 in Fig. 1).

As a result, we expect both overall growth effects, b_{10} and b_{20} , to be positive. However, an individual student might show a smaller or no growth through the course—for instance, if he/she knows the content before the course, or, on the contrary, if the course is too difficult for the student to understand. In this case, the student's random effect b_{1j} would be negative: The learning rate would be smaller than the overall learning rate across students. As we mentioned before, it is also possible that a student could simply enumerate possibilities, for example by clicking repeatedly on the alternative options in a multiple-choice question. This would mean there would be no real growth within the learning objective. Therefore, the effect of attempt for this student would be smaller than for an average student, and hence his/her deviation b_{2j} from the mean effect b_{20} would be expected to be negative. For a student who was learning fast both throughout the course and within a certain learning objective, in contrast, we would expect both the student-specific random effects b_{1j} and b_{2j} to be positive. Note that this approach can be distinguished from modeling possible guessing with a three-parameter logistic model (3PL; Lord & Novick, 1968). First, the 3PL model is used for items that are attempted only once. It models pseudo-guessing, but not a growth in probability that comes with extra attempts. Second, the 3PL model is applicable to multiple-choice items only, whereas the proposed extension might deal with open-ended items as well, which are widely used in MOOCs.

In the following section, we illustrate these extensions using the data from three MOOCs and check the performance of the extensions using a cross-validation procedure applied to the same datasets.

Method

Data

In this study, we used data from three MOOCs on the Coursera platform: “Economics for Non-Economists” (Higher School of Economics, n.d.-a), “Game Theory” (Higher School of Economics, n.d.-b), and “Introduction to Neuroeconomics: How the Brain Makes Decisions” (Higher School of Economics, n.d.-c). We analyzed the data from five weekly modules for each course.

In the data, each student and all course elements, such as a video lecture or an assessment item, has a unique identification number. Each interaction of students with course elements also has an individual identification number and a time stamp. Students' responses on summative assessment items have a dichotomous coding, where 1 and 0 correspond to the correct and a wrong response on a certain attempt, respectively. The assessment items are either multiple-choice or open-ended questions, in which a student is expected to choose one or multiple options or to respond with a number,

sequence, word, or phrase. There is no overlap in items between the summative assessments in different weeks, and the correctness of students' responses is checked automatically. Attempts are marked with a unique time stamp. Student's interactions with video lectures are coded as 0 or 1, where 1 means the student watched the lecture and 0 means the student did not watch the lecture. The platform does not track how many times a student watches a certain video.

The first course, "Economics for Non-Economists," is taught in Russian. At the moment of conducting this study, there were 1,632 active students in the course. The distribution of students among countries was as follows: Russia (72%), Ukraine (8.4%), Kazakhstan (3.9%), Belarus (3.2%), USA (1.2%), and Other (11.3%). The number of items in the weekly summative assessments for these five modules was 68 in total: ten items for Weeks 1 to 4, eight items for Week 5, and 20 items in a concluding assessment. The total number of responses for the first course was 134,068. Students used 1.89 attempts on average, with a standard deviation of 1.41. After recoding the attempts to 0, 1, 2, 3, and 4—which mean the first, second, third, fourth, and fifth or higher attempts, respectively—the mean of attempts was 0.82 and the standard deviation 1.12. The number of video lectures for these five modules was 48 in total: nine in the first week, and eight, nine, 13, and nine in the following weeks, respectively. Students watched on average 6.70, 6.60, 7.35, 9.99, and 7.27 videos, with standard deviations of 3.00, 2.30, 2.65, 4.32, and 2.79 during the first and following weeks, respectively.

The second course, "Game Theory," is taught in Russian. The distribution of students among countries was as follows: Russia (57%), Ukraine (10%), Kazakhstan (3.3%), USA (3.2%), Belarus (3%), and Other (23.5%). The third course, "Introduction to Neuroeconomics: How the Brain Makes Decisions," is taught in English. The distribution of students among countries was as follows: USA (19%), India (8.7%), Russia (6.8%), Mexico (4.7%), United Kingdom (4%), and Other (56.8%).

Descriptive statistics for the students, items, responses, attempts, and video lectures for these courses are given in Table 1.

Illustration of the extensions

We start with the Rasch model from Eq. 5. As we discussed above, the Rasch model has no growth effects. However, we fitted this model in order to get a benchmark for comparing the model and its dynamic extensions, and to provide an empirical check for the unidimensionality assumption. Then we continued by fitting the extension with fixed growth effects from Eq. 6 and the extension with (uncorrelated and correlated) random growth effects from Eq. 7. To fit the Rasch model and the extensions, we used the glmer function in the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) for the R language

Table 1 Data overview

	Course 1	Course 2	Course 3
Students	1,632	3,092	4,873
Items	68	50	60
Responses	134,068	228,490	339,330
Attempts	1.89 (1.41)	1.92 (1.50)	2.10 (2.01)
Video Lectures	48	44	26
Numbers of video lectures looked at in:			
Week 1	6.70 (3.00)	6.14 (2.66)	3.75 (1.57)
Week 2	6.60 (2.30)	7.53 (2.63)	3.18 (1.15)
Week 3	7.35 (2.65)	8.31 (2.96)	5.26 (2.24)
Week 4	9.99 (4.32)	6.98 (2.10)	5.12 (1.62)
Week 5	7.27 (2.79)	7.65 (2.54)	3.63 (0.89)

In the table, Course 1 is "Economics for Non-Economists" (Higher School of Economics, [n.d.-a](#)), Course 2 is "Game Theory" (Higher School of Economics, [n.d.-b](#)), and Course 3 is "Introduction to Neuroeconomics: How the Brain Makes Decisions" (Higher School of Economics, [n.d.-c](#)). For Attempts and Numbers of Video Lectures, we show mean values and, in parentheses, standard deviations

and environment for statistical computing (R Core Team, 2013). To check the unidimensionality assumption, we used the unidimTest function in the ltm package (Rizopoulos, 2006) for R, which implements the approach proposed by Drasgow and Lissak (1983), in which the latent dimensionality is checked via a comparison of the eigenvalues from a factor analysis of the observed data and from data generated under the assumed unidimensional IRT model. The null hypothesis of unidimensionality is rejected if the second eigenvalue is substantially larger for the observed than for the simulated data. To approximate the distribution of the test statistic under the null hypothesis, we used 100 samples in the Monte Carlo procedure implemented by the unidimTest function. To compare the model fits, we used the Akaike information criterion (AIC; Akaike, 1974) provided by the glmer function.

Cross-validation

When we talk about the value of the extensions in the cross-validation, we refer to its accuracy at predicting the correctness of students' responses in the summative assessments (Ekanadham & Karklin, 2015). The procedure looks as follows: First we randomly split the students' responses on a summative assessment from the three MOOCs into training and test datasets. We assigned 75% of all responses to the training set and 25% to the test set. Then we fitted the Rasch model and both extensions to the training sets and used the fitted models to derive the probabilities of correct responses on the test sets. Using the probabilities, we derived the expected responses, in the sense that if the probability was smaller than .50, the expected response was 0, and if it was larger than .50, the expected response was 1. Next, using the predicted and real responses

Table 2 Parameters of the extension for Course 1

		Rasch model (Eq. 5)	Extension with fixed growth effects (Eq. 6)	Extension with random growth effects and univariate distribution (Eq. 7)	Extension with correlated random growth effects and multivariate distribution (Eq. 7)	
Fixed	Intercept b_0	0.96 (0.09)	– 0.52 (0.16)	– 0.38 (0.17)	– 0.32 (0.17)	
	Video b_1		4.45 (0.45)	3.71 (0.50)	3.72 (0.46)	
	Attempt b_2		0.43 (0.01)	0.80 (0.02)	0.82 (0.02)	
Random	Student Intercept σ_{u1}	0.72	0.79	0.80	0.95	Corr.
	Video σ_{b1}			1.52	2.07	– .67
	Attempt σ_{b2}			0.52	0.51	.14
Item	Intercept σ_{i2}	1.03	1.09	1.09	1.08	.02
	AIC	146,979	143,024	140,367	140,221	

In the table, Course 1 is “Economics for Non-Economists” (Higher School of Economics, [n.d.-a](#)). For fixed effects, standard errors are presented in parentheses. For random effects, standard deviations are presented

from the test sets, we built a confusion matrix with two rows and two columns, for the numbers of false positives, false negatives, true positives, and true negatives, respectively. This matrix allowed us to understand the quality of the predictions as well as the improvements in prediction across the models. Using this matrix, we calculated an accuracy index:

$$Accuracy = \frac{TP + TN}{P + N}, \quad (8)$$

where TP is true positives; TN is true negatives; P is all positives; and N is all negatives. Then we repeated this procedure five times and finished by counting the average accuracy for each model.

Results

We start from the Rasch model from Eq. 5. The dimensionality check did not reveal significant evidence against unidimensionality ($p = .15$). As is shown in Table 2, the estimate of the intercept equals 0.96. The inverse logit, or *antilogit*, of this value is 0.72. This means that the expected probability that an average student of the “Economics for Non-Economists” course would give the correct response on an average item

was .72. This probability of the correct response would vary among students (and over items). A student with proficiency of one standard deviation lower and a student with proficiency of one standard deviation higher than the average proficiency would have proficiencies of 0.24 and 1.68, which correspond to probabilities of a correct answer on an average item of .56 and .84, respectively. As we mentioned above, the Rasch model assumes that the student’s proficiency remains constant within the course.

However, extending the Rasch model with fixed growth effects (the first extension; Eq. 6), shows that the probability of a correct response grew with every new watched lecture and with every new attempt to solve an item. As can be derived from Table 2, watching all nine video lectures from the first week would improve the chances of a correct response for an average student on an average item from .37 to .47 (see Table 3); these are the antilogits of -0.52 and $(-0.52 + 9/100 \cdot 4.45)$. By the end of the third week, the chances of the correct response (on the same average item) would rise to .65. By the end of the fifth week, the chances of the correct response would approach the value of .83 (Table 3). To get a better idea of the size of the growth within a certain learning objective, we compared the chances of a correct response for an average student on an average item from the first, second, and third attempts. These values were .37, .48, and .58,

Table 3 Dynamics of antilogits throughout Course 1

	Start	Week 1			Week 5		
	Avg.	Avg.	– SD	+ SD	Avg.	– SD	+ SD
Rasch model (Eq. 5)	.72	.72			.72		
Extension with fixed growth effects (Eq. 6)	.37	.47			.83		
Extension with random growth effects and univariate distribution (Eq. 7)	.41	.49	.45	.52	.80	.66	.89
Extension with correlated random growth effects and multivariate distribution (Eq. 7)	.42	.50	.46	.55	.81	.62	.92

In the table, antilogits for a student with average (initial) ability and for an item with average difficulty are presented. The dynamics of antilogits for the start, the end of the first week, and the end of the fifth week of the course are presented. For the model with random growth effects, the average, one standard deviation lower than the average, and one standard deviation higher than the average continuous growth effect are presented. Course 1 is “Economics for Non-Economists” (Higher School of Economics, [n.d.-a](#))

Table 4 Dynamics of antilogits with attempts in Course 1

	Att. 1	Attempt 2			Attempt 3		
	Avg.	Avg.	– SD	+ SD	Avg.	– SD	+ SD
Rasch model (Eq. 5)	.72	.72			.72		
Extension with fixed growth effects (Eq. 6)	.37	.48			.58		
Extension with random growth effects and univariate distribution (Eq. 7)	.41	.60	.48	.72	.77	.54	.91
Extension with correlated random growth effects and multivariate distribution (Eq. 7)	.42	.62	.50	.73	.79	.57	.91

In the table, antilogits for a student with average ability and for an item with average difficulty at the start of the course are presented. The dynamics of antilogits for the first, second, and third attempts are presented. For the model with random growth effects, the average, one standard deviation lower than the average, and one standard deviation higher than the average local growth effect are presented. Course 1 is “Economics for Non-Economists” (Higher School of Economics, [n.d.-a](#))

respectively (Table 4). It is important to note that, in accordance with the extension, the growth in both components was assumed to be the same for all students.

If we allow the growth parameters to vary over students in accordance with the second extension (Eq. 7), we can derive the individual differences in both components. First, we look at the results of the first version of the second extension, with univariate distributions of random effects. As can be seen in Table 3, by the end of the first week two students with an average initial proficiency but growth-through-the-course parameters one standard deviation lower and one standard deviation higher than the average would have probabilities of a correct answer of .45 and .52, respectively. By the end of the fifth week, these values would be .66 and .89 for the two students, respectively.

The second component of the growth in proficiency, the growth within a certain learning objective, varies among students as well. As can be found in Table 4, on average the chances of a correct response grow from .41 for the first attempt, over .60 for the second attempt, to .77 for the third attempt. However, two students with average initial proficiency but growth effect parameters within a certain learning objective of one standard deviation lower and one standard deviation higher than the average would, at the second attempt,

have probabilities of a correct answer of .48 and .72, respectively. For the same students, the chances of a correct response on the third attempt would be .54 and .91, respectively.

The results for the second version of the second extension (Eq. 7), with a multivariate distribution of random effects, presented in Table 2, allow us to understand the relations between student-specific effects. The moderate negative correlation between the student-specific random intercept and the random effect of video lectures (–.67) advises us that students with lower initial proficiency would show a higher effect of the number of watched video lectures on their performance in the course. At the same time, there are only weak correlations between the student-specific random intercept and the random effect of attempts and between both random slopes (.14 and .02, respectively).

In the other two courses, the dimensionality check again did not show substantial evidence against unidimensionality (with p values equal to .30 and .07). We detected similar growth through the course and within a certain learning objective in both these courses. As can be derived from Tables 5 and 6, in the “Game Theory” and “Introduction to Neuroeconomics: How the Brain Makes Decisions” courses, the probabilities of correct response grow with every new watched lecture and with every new attempt to solve an item.

Table 5 Parameters of the extension for Course 2

		Rasch model (Eq. 5)	Extension with fixed growth effects (Eq. 6)	Extension with random growth effects and univariate distribution (Eq. 7)	Extension with correlated random growth effects and multivariate distribution (Eq. 7)
Fixed	Intercept b_0	0.91 (0.06)	0.08 (0.11)	0.01 (0.11)	0.00 (0.11)
	Video b_1		2.67 (0.35)	2.86 (0.38)	3.01 (0.39)
	Attempt b_2		0.68 (0.01)	1.02 (0.02)	1.05 (0.02)
Random Student	Intercept σ_{u1}	0.67	0.82	0.69	0.72
	Video σ_{b1}			2.44	2.69
	Attempt σ_{b2}			0.50	0.50
Item	Intercept σ_{u2}	0.67	0.77	0.79	0.79
		262,254	247,160	243,615	243,394

In the table, Course 2 is “Game Theory” (Higher School of Economics, [n.d.-b](#)). For fixed effects, the standard error is presented in parentheses. For random effects the standard deviation is presented

Table 6 Parameters of the extension for Course 3

		Rasch model (Eq. 5)	Extension with fixed growth effects (Eq. 6)	Extension with random growth effects and univariate distribution (Eq. 7)	Extension with correlated random growth effects and multivariate distribution (Eq. 7)		
Fixed	Intercept b_0	1.56 (0.15)	0.55 (0.18)	0.30 (0.18)	0.28 (0.20)		
	Video b_1		5.53 (0.57)	7.05 (0.59)	7.28 (0.82)		
	Attempt b_2		0.39 (0.01)	0.55 (0.01)	0.57 (0.01)		
Random Student	Intercept σ_{u1}	0.79	0.87	0.86	1.02	Corr.	
	Video σ_{b1}			2.84	3.80	– .62	
	Attempt σ_{b2}			0.32	0.32	.06	.06
Item	Intercept σ_{i2}	1.27	1.35	1.41	1.41		
	AIC	293,177	287,006	284,775	284,512		

In the table, Course 3 is “Introduction to Neuroeconomics: How the Brain Makes Decisions” (Higher School of Economics, [n.d.-c](#)). For fixed effects, the standard error is presented in parentheses. For random effects the standard deviation is presented

Both growth parameters vary among students and show individual differences in both components.

Value of the extensions

The model fit improved with each extension. Table 2 shows that for the “Economics for Non-Economists” course, the AIC decreases from 146,979 for the Rasch model to 143,024 for the extension with fixed growth effects, and then to 140,367 and 140,221 for the extensions with uncorrelated and correlated random growth effects, respectively. Similar improvements of the model fit were found for the other two courses (Tables 5 and 6).

The results of the cross-validation procedure, presented in Table 7, show that over the three courses, both extensions are better at measuring proficiency and its growth than the Rasch model, in terms of accuracy at predicting students’ responses. The overall accuracy of the Rasch model is .743, whereas the overall accuracies of the extensions are .760 for the extension with fixed growth effects, and .766 for each of the extensions with uncorrelated and correlated random growth effects. We see such improvements in accuracy in all three courses we analyzed. The improvements are rather small, however, and they come with higher numbers of model parameters.

Discussion and conclusion

In this study, two extensions for the Rasch model for measuring the growth in students’ proficiency in MOOCs were presented. First, the study has contributed to psychometric methodology. It focused on existing ideas of modeling dynamic processes in the framework of IRT but extended their ability to detect the novel latent growth trends that appear in datasets from MOOCs. For instance, where Verhelst and Glas (1993) and De Boeck et al. (2011) presented two ways to generalize the traditional Rasch psychometric model by making it dynamic, in our extensions we built on this idea by modeling two components of dynamic processes in students’ proficiency—continuous and local growth. Second, we introduced IRT in a new application area. Whereas Kadengye, Ceulemans, and Van den Noortgate (2014, 2015) applied their models to item-based learning environments, we implemented IRT models in the new and fast-developing context of MOOCs. Finally, our findings are important for the practice of online learning. Measuring growth in students’ proficiency might be essential for MOOC developers, because these measures can help us understand how efficient a course is in terms of the group and individual dynamics of students, and for students themselves to understand their personal progress, in the sense of formative feedback.

Table 7 Accuracy in predicting correctness

	Overall		Course 1		Course 2		Course 3	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Rasch model	.743	.047	.724	.002	.699	.002	.806	.000
Extension with fixed growth effects	.760	.038	.737	.002	.732	.002	.812	.001
Extension with random growth effects and univariate distribution	.766	.034	.747	.002	.740	.001	.813	.001
Extension with random growth effects and multivariate distribution	.766	.034	.747	.002	.739	.001	.812	.001

In the table, Course 1 is “Economics for Non-Economists” (Higher School of Economics, [n.d.-a](#)), Course 2 is “Game Theory” (Higher School of Economics, [n.d.-b](#)), and Course 3 is “Introduction to Neuroeconomics: How the Brain Makes Decisions” (Higher School of Economics, [n.d.-c](#))

The study, however, has limitations. We used the proposed extensions to measure growth post-hoc, not to track growth dynamically, which will be crucial for developing navigation and recommendation instruments that can help teachers decide when to support a student or advance him or her through a course, by means of on-the-fly estimations of progress. We believe this goal could be realized by combining these models with techniques for tracking growth—for example, with the Elo (1978) rating system, used in the work of Klinkenberg, Straatemeier, and van der Maas (2011). This will be a topic for future research.

To conclude, we consider this study an additional step in transition from traditional psychometric approaches, focused on accurately locating students on a proficiency scale, to flexible approaches based on computational psychometrics (von Davier, 2017), oriented toward regarding students' behavior as a dynamic process.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, 50, 3–16.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Coursera. (n.d.). Producing engaging video lectures. Retrieved from Coursera Partner Resource Center: <https://partner.coursera.help/hc/en-us/articles/203525739-Producing-Engaging-Video-Lectures>
- Davis, F. B. (1964). *Educational measurements and their interpretation*. Belmont, CA: Wadsworth.
- De Boeck, P., Bakker, M., Zwister, 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.
- Drasgow, F., & Lissak, R. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, 68, 363–373.
- Ekanadham, C., & Karklin, Y. (2015, July). *T-SKIRT: Online estimation of student proficiency in an adaptive learning system*. Paper presented at the 31st International Conference on Machine Learning, Lille, France.
- Elo, A. (1978). *The rating of chessplayers, past and present*. New York, NY: Arco.
- Fisher, G. H. (1976). Some probabilistic models for measuring change. In D. N. De Gruijter, & L. J. Van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97–110). New York, NY: Wiley.
- Fisher, G. H. (1995). Linear logistic models for change. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 157–180). New York, NY: Springer.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Higher School of Economics. (n.d.-a). Economics for non-economists. Retrieved from Coursera: <https://www.coursera.org/learn/ekonomika-dlya-neekonomistov>
- Higher School of Economics. (n.d.-b). Game theory. Retrieved from Coursera: <https://www.coursera.org/learn/game-theory>
- Higher School of Economics. (n.d.-c). Introduction to neuroeconomics: How the brain makes decisions. Retrieved from Coursera: <https://www.coursera.org/learn/neuroeconomics>
- 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.
- Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. (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.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.
- Molenaar, I. W. (1995). Some background for Item Response Theory and the Rasch model. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 3–14). New York, NY: Springer.
- R Core Team. (2013). R: A language and environment for statistical computing (R Foundation for Statistical Computing) Retrieved from <http://www.R-project.org/>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17, 1–25.
- Shah, D. (2016). Monetization over massiveness: Breaking down MOOCs by the numbers in 2016. Retrieved from EdSurge: <https://www.edsurge.com/news/2016-12-29-monetization-over-massiveness-breaking-down-moocs-by-the-numbers-in-2016>
- 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.
- Verguts, T., & De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement*, 24, 151–162.
- Verhelst, N. D., & Glas, C. A. (1993). A dynamic generalization of the Rasch model. *Psychometrika*, 58, 395–415.
- Verhelst, N. D., & Glas, C. A. (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.
- von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, 54, 3–11.