Measuring student’s proficiency in MOOCs: multiple attempts extensions for the Rasch model

Dmitry Abbakumov a,b,c,* , Piet Desmet a,d, Wim Van den Noortgate a,b

a ITEC imec, Leuven, Belgium
b Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium
c eLearning Office, National Research University Higher School of Economics, Moscow, Russia
d Faculty of Arts, KU Leuven, Leuven, Belgium

* Corresponding author.
E-mail address: dmitry.abbakumov@kuleuven.be (D. Abbakumov).

Abstract

Popularity of online courses with open access and unlimited student participation, the so-called massive open online courses (MOOCs), has been growing intensively. Students, professors, and universities have an interest in accurate measures of students’ proficiency in MOOCs. However, these measurements face several challenges: (a) assessments are dynamic: items can be added, removed or replaced by a course author at any time; (b) students may be allowed to make several attempts within one assessment; (c) assessments may include an insufficient number of items for accurate individual-level conclusions. Therefore, common psychometric models and techniques of Classical Test Theory (CTT) and Item Response Theory (IRT) do not serve perfectly to measure proficiency. In this study we try to cover this gap and propose cross-classification multilevel logistic extensions of the common IRT model, the Rasch model, aimed at improving the assessment of the student’s proficiency by modeling the effect of attempts and by involving non-assessment data such as student’s interaction with video lectures and practical tasks. We illustrate these extensions on the logged data from one MOOC and check the quality using a cross-validation procedure.
on three MOOCs. We found that (a) the performance changes over attempts depend on the student: whereas for some students performance ameliorates, for other students, the performance might deteriorate; (b) similarly, the change over attempts varies over items; (c) student’s activity with video lectures and practical tasks are significant predictors of response correctness in a sense of higher activity leads to higher chances of a correct response; (d) overall accuracy of prediction of student’s item responses using the extensions is 6% higher than using the traditional Rasch model. In sum, our results show that the approach is an improvement in assessment procedures in MOOCs and could serve as an additional source for accurate conclusions on student’s proficiency.

Keywords: Education, Psychology

1. Introduction

Massive open online courses (MOOCs) are a recent progressive phenomenon in education. A MOOC is an online course with open access and participation of unlimited number of students. A MOOC typically consists of pre-recorded video lectures, reading assignments, assessments, and forums. MOOCs are mainly developed by universities and run on platforms such as Coursera, edX, XuetangX, FutureLearn, Udacity, MiriadaX. In 2017, more than 800 universities offered students more than 9,400 MOOCs (Shah, 2018). The same year, the largest MOOC provider, Coursera, achieved the milestones of 30 million students and 2,700 courses (Shah, 2017).

Students, professors and universities — the key partners involved in MOOCs, — have an interest in accurate student proficiency measuring. Students take an online course and want to study efficiently. Proficiency measuring specifies student’s position on the course-line, helps him/her to identify his/her strong and weak points and map areas that need additional work. Professors and their teams develop and optimize the course content. At this stage, the aggregated proficiency measures show to what degree the content incites learning and suggest improvements of video lectures, practical tasks, and support materials. Finally, universities award online course certificates to students. Proficiency measures can provide evidence whether student has mastered the course.

Proficiency is, however, a latent variable. We cannot observe it directly; yet, we can see, for example, student’s performance on assessment items. To link the observable side to the latent side, we need specific rules. These rules are called measurement or psychometric theory. Today, there exist two major psychometric theories: Classical Test Theory (CTT) and Item Response Theory (IRT) and dozens of models based on them.
CTT appeared around 100 years ago (Borsboom, 2005). It includes three main concepts — test score (observed), true score (latent), and error score (latent) (Lord and Novick, 1968). The theory introduces a simple linear model that links the observable test score to the sum of two latent variables, true score and error score, that is,

\[ Y_j = \theta_j + \varepsilon_j. \]  

(1)

According to this model, the test score of the \( j \)'th student \( (Y_j) \) is the result of his/her proficiency \( (\theta_j) \) with a random measurement error \( (\varepsilon_j) \). The error term \( (\varepsilon_j) \) has the expected value of zero, is assumed normally distributed and unrelated to the proficiency: \( E(\varepsilon_j) = 0, \varepsilon_j \sim N(0, \sigma^2_\varepsilon) \), and \( \rho_{\varepsilon\theta} = 0 \). Thus, the expected value of \( Y_j \), \( E(Y_j) \), is \( \theta_j \). As a result, the average score is normally distributed around \( \theta_j \) with variance \( \sigma^2_\varepsilon/n \) with \( n \) being the number of observations. Hence, the more observations, the closer the average score is in general to the proficiency.

The CTT has several limitations (see Hambleton and Jones, 1993). For MOOCs the critical limitation is the dependence of proficiency measures on test difficulty: when the test is difficult, the student receives a low estimate of proficiency and when the test is easy, the student receives a high estimate of proficiency. This is a problem for assessments in MOOCs, since changes in tests are relatively frequent in MOOCs — professors often replace or add new items on the fly. Universities run the risk of making unfair decisions in certification of students. Professors will receive biased information about learning materials functioning. Students will have the wrong map of their strengths and weaknesses.

IRT was developed 50 years ago (Lord and Novick, 1968; Rasch, 1960). It presents a class of models with nonlinear linking between student’s item responses (observable variables) and his/her proficiency (latent variable). In the basic IRT model, the Rasch model (Rasch, 1960),

\[ \logit(\pi_{ij}|\theta_j) = \ln(\pi_{ij}/1-\pi_{ij}) = \theta_j - \delta_i \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}), \]  

(2)

probability \( (\pi_{ij}) \) of the correct response of student \( j \) to the item \( i \) is described by a logistic function of the difference between the student’s proficiency parameter \( (\theta_j) \) and the item difficulty parameter \( (\delta_i) \). To fit the Rasch model, a marginal maximum likelihood procedure is often used to estimate the item difficulty parameters, assuming that the students are a random sample from population where the student proficiencies are normally distributed with \( \theta_j \sim N(0, \sigma^2_\theta) \), while the items have fixed difficulty. Individual student parameters can be estimated afterwards using empirical Bayes procedures.

In comparison to CTT, student’s proficiency parameters in IRT are independent from the test difficulty. It allows comparing students even in case of partial replacement of items. However, the use of IRT in MOOCs is also challenging. Firstly, IRT requires a relatively large number of items in assessments to provide accurate
proficiency measures. Kruyen et al. (2012) stated that the minimal test length for individual level decisions is 40 items. By contrast, MOOCs often offer 15 or even fewer items in summative assessments. At the same time, MOOCs yield additional observable data such as student’s activity with video lectures and his/her performance in practice, which might be used as proficiency indicators to cover the lack of items in summative assessments. Secondly, the proficiency parameter in IRT as well as in CTT is typically considered constant (Lord and Novick, 1968). This point contrasts to the reality of MOOCs where students get feedback after responding to assessment items. In addition, students may be allowed to do several attempts within one assessment. If the student fails at one attempt, he/she can be provided with help information, review a video lecture, and then make a new attempt. Thus, the student’s proficiency may grow with each new attempt to solve the certain item. However, better performance with new attempts does not necessarily imply growth in proficiency, as it might also appear if the student uses attempts to enumerate the item options. The challenge here is to distinguish between these kinds of growth.

As can be seen from the above, the common psychometric models of CTT and IRT are not tailored to use directly for measuring proficiency in MOOCs. At the same time, IRT is a well-elaborated and flexible framework, and could serve as the basis for such measurements.

In this study we extend and tune up the common IRT model, the Rasch model, for application in MOOCs. We start with proposing four extensions, which model the growth of student’s proficiency with attempts and involve data of student’s interaction with video lectures and practical tasks to compensate for the insufficient number of items in the assessment. Then we illustrate these extensions using data from one MOOC. Finally, we check the performance of these extensions in predicting correctness of students’ responses in summative assessments and show the advantage in accuracy in comparison to the common IRT model using a cross-validation procedure applied in the data from three MOOCs.

2. Model

As the general framework, we use reformulation of the Rasch model presented by Van den Noortgate et al. (2003):

\[
\text{Logit}(\pi_{ij}) = b_0 + u_{1j} + u_{2j} \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}),
\]

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

This reformulation is based on the principle of cross-classification multilevel models. Instead of two parameters (student’s proficiency parameter and item difficulty parameter), in this reformulation, we have an intercept and two residual terms referring to the student and the item respectively. The intercept equals the estimated
logit of probability of a correct response of an average student on an average item. The intercept can be interpreted as the difference between the overall proficiency and the overall difficulty. The mean of both residual terms is zero. The first residual term shows the deviation of proficiency of student \( j \) from the overall proficiency. The second residual term shows the deviation of the difficulty of item \( i \) from the overall difficulty, in the sense that the larger the residual, the easier the item.

We prefer this reformulation to the traditional Rasch model formulation because it is very flexible for making extensions. Considering items and students as random, there are still degrees of freedom left to include various predictors in the model, whereas, for instance, inserting item characteristics as predictors to a model where the items are considered as fixed (and therefore the difficulty is estimated for each item separately) leads to computational issues because of overidentification of the model (Van den Noortgate et al., 2003).

Before presenting extensions, we should describe the general structure of MOOCs. In this study we focus on the structure of courses on Coursera, but this structure is common for other MOOCs. Courses on Coursera are composed of modules, which consist of lessons. Each lesson includes several video lectures accompanied by at least one task of formative assessment. Each video lecture lasts 4–9 minutes. Thus, it takes a student about 30 minutes to complete a lesson. Typically, each lesson is structured around one or two objectives of learning. The module is a collection of lessons that comprises a larger unit of learning. Each module is structured around a cohesive subtopic and usually lasts a week. Modules are concluded by summative assessment. Summative assessments are realized as a 10–15 item test, a programming task, or a peer-review task. In this study we use courses with test-based summative assessments; we work with the data of each weekly summative assessment separately.

2.1. Extension one

In the first extension we consider two components of proficiency within weekly summative assessment: the fixed component and the dynamic component. We assume that during the assessment the student’s proficiency is fixed. However, if the student takes several attempts, there may be a change in the expected student’s performance from one attempt to another within the assessment. To model this growth, we use the dynamic component.

Thus, in our first extension,

\[
\text{Logit}(\pi_{ij}) = b_0 + (b_{10} + b_{1j}) \cdot \text{attempt}_{ij} + u_{ij} + u_{2i} \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}), \tag{4}
\]

\( b_0 \) equals the estimated logit of probability of the correct response of an average student on an average item in weekly summative assessment; \( \text{attempt}_{ij} \) is 0, 1, 2,
3 or 4 and means the first, the second, the third, the fourth, or the fifth or higher attempt respectively; $b_{10}$ is overall effect of attempt, while $b_{1j} \sim N(0, \sigma_{b1}^2)$ is a deviation of the attempt effect for student $j$ from the overall effect; and $u_{1j} \sim N(0, \sigma_{u1}^2)$ and $u_{2i} \sim N(0, \sigma_{u2}^2)$.

We expect a positive average effect of attempts ($b_{10}$). It means that with each new attempt the chances for the correct response grow. The first reason for this is learning. For instance, online courses provide students with hints in case of a wrong response. This instructional content is aimed at helping the student to understand his/her mistake, guiding through relevant learning materials and preparing for the next attempt. The second reason is that the student can simply enumerate possibilities, for example, by clicking repeatedly on alternative options in multiple-choice questions. Therefore we expect that increase of the correct response probability for a student who does not learn at all, will be lower than with the average student, that is, the student’s random effect $b_{1j}$ is expected to be negative, and its combination with the average effect of attempts ($b_{10}$) is expected to be close to zero. For fast learning students, on the contrary, we expect positive random effect $b_{1j}$.

2.2. Extension two

It is important to note that the effect of additional attempt may vary from item to item. For some items, the expected increase of probability of giving the correct response from one attempt to the next attempt may be larger than for other items. For instance, a multiple-choice item with four options could be solved correctly (using simple enumeration) by four attempts maximally. For solving an open-ended item, where the student should indicate a number or a word, the number of attempts may be much higher. To account for this variation, we propose the second extension:

$$\text{Logit}(\pi_{ij}) = b_0 + (b_{10} + b_{1j} + b_{1i}) \ast \text{attempt}_{ij} + u_{1j} + u_{2i} \land Y_{ij} \sim \text{Bernoulli}(\pi_{ij}),$$

where $b_{1i} \sim N(0, \sigma_{b1}^2)$ is a deviation of the attempt effect for item $i$ from the overall effect.

2.3. Extension three

As students with different patterns of using attempts may have different growth with each new attempt, we propose the third extension. We could split the course students in classes by maximum number of attempts, they used to solve items in weekly summative assessment, where for instance, classes “0”, “1”, “2”, and “3” mean that the maximum number of attempts for this student was 1 or 2, 3 or 4, 5 or 6, and more than 6 respectively. We use these classes in interaction with the number of attempts in the third extension:
\[
\text{Logit}(\pi_{ij}) = b_0 + (b_{10} + b_{11} + b_{12}) \cdot \text{attempt}_j + b_2 \cdot \text{class}_j + b_3 \cdot \text{attempt}_j \cdot \text{class}_j \\
+ u_{1j} + u_{2i} \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}),
\]

where \( \text{class}_j \) is 0, 1, 2, and 3.

We expect that the more attempts one makes, the less he/she learns per an additional attempt, and hence, \( b_3 \) is negative. We also expect a negative main effect of the class, in the sense that students with a higher maximum number of attempts are less proficient. For instance, it could be students, interested in formal achievements in assessment rather than in mastering the course.

2.4. Extension four

Finally, we propose to include predictors indicating student’s week activity before assessment, for instance, how active he/she was in watching video lectures and how productive he/she was with formative assessments. These predictors may show systematic differences in probability of the correct response for students with different quality of course activity. Thus, our fourth extension is:

\[
\text{Logit}(\pi_{ij}) = b_0 + (b_{10} + b_{11} + b_{12}) \cdot \text{attempt}_j + b_2 \cdot \text{class}_j + b_3 \cdot \text{attempt}_j \cdot \text{class}_j \\
+ b_4 \cdot \text{formative.assessment.performance}_j + b_5 \cdot \text{lecture.activity}_j + u_{1j} \\
+ u_{2i} \text{ and } Y_{ij} \sim \text{Bernoulli}(\pi_{ij}),
\]

where \( \text{formative.assessment.performance}_j \) is the index that refers to the student’s performance in the previous formative assessments of the module; it equals the number of correct attempts, divided by the sum of the number of student’s correct and wrong attempts, that may take values from 0 to 1; \( \text{lecture.activity}_j \) is the index that refers to student’s activity with video lectures in the module; it takes values from 0, which means that the student did not watch video lectures at all, to 2, which means that the student watched all video lectures in the module. Therefore, \( b_4 \) and \( b_5 \) are overall effects of performance in formative assessments and activity with video lectures respectively. The coefficients of the formative assessment performance and the lecture activity do not vary from student to student because these variables were measured only once, before the summative assessment starts, and therefore only one value of the variable corresponds to each single student. In this extension we expect these both effects to be positive.

3. Methods

The course data we use are initially anonymized and coded in the following way. Students have a unique identification number. The course and all its elements — a
module, a lesson, a video lecture, and an item in assessment, — have their individual identification number. Student’s actions, for instance interaction with video lecture or item response, also have an individual identification number and a time stamp. Student’s responses on formative and summative assessment items are dichotomous variables where 1 and 0 correspond to correct and wrong response in a certain attempt respectively. Attempts are marked with the time stamp. Student’s interactions with each video lecture are coded as 0, 1 or 2 where 2 means the student finished watching the lecture, 1 means the student started but did not finish watching the lecture, 0 means the student did not start watching the lecture. The platform does not track intermediate progress stages, which means, for instance, that 70% and 99% of a progress in watching a video lecture are coded in the same way (using code 1). We use the last record for each video, which indicates the actual state of the student with certain video during weekly assessment.

To illustrate the extensions, we fit them on the data of “Economics for Non-Economists” MOOC on Coursera (Higher School of Economics, n.d.). This course is taught in Russian. The distribution of students among countries is as follows: Russia (72%), Ukraine (8.4%), Kazakhstan (3.9%), Belarus (3.2%), USA (1.2%), other countries (11.3%). The percentage of female students is 47%. We used the data from the first module of the course. During this study, the number of students who attended the module was 1609. The weekly summative assessment includes 10 items. The number of responses is 51,550. Students used 2.04 attempts in average, the standard deviation is 1.52. After recoding the attempts to 0, 1, 2, 3, and 4 that means the first, the second, the third, the fourth, or the fifth or higher attempt respectively, the mean of attempts is 0.95 and standard deviation is 1.19. The average student’s activity with video lectures is 1.87 and standard deviation is 0.38, where the minimum value of 0 means that the student did not watch any lectures at all, and the maximum value of 2 means that the student watched all lectures till the end. The average student performance with the formative assessment during the week is .60 and standard deviation is .28, where the minimum value of 0 means that the number of his/her correct attempts was 0, and the maximum value of 1 means that all his/her first attempts were correct (or the number of his/her correct attempts, divided by the sum of the number of correct and wrong attempts was 1).

To show the performance of the extensions in the cross-validation procedure we used three MOOCs on Coursera: “Economics for Non-Economists”, “Game Theory” (Higher School of Economics, n.d.), and “Introduction to Neuroeconomics: How the Brain Makes Decisions” (Higher School of Economics, n.d.). The “Game Theory” course is taught in Russian. The distribution of students among countries is as follows: Russia (57%), Ukraine (10%), Kazakhstan (3.3%), USA (3.2%), Belarus (3%), other countries (23.5%). The percentage of female students is 37%. The “Introduction to Neuroeconomics: How the brain Makes Decisions” course is taught in English. The distribution of students among countries is as
follows: USA (19%), India (8.7%), Russia (6.8%), Mexico (4.7%), United Kingdom (4%), other countries (56.8%). The percentage of female students is 41%. We used the data from first three modules of each of these courses. Selected descriptive statistics of these courses are presented in Table 1.

To fit the extensions, the “glmer” function of the “lme4” package (Bates et al., 2015) of the R language and environment for statistical computing (R Core Team, 2013) can be used. The “lme4” package is used for fitting various IRT models (De Boeck et al., 2011). To compare the models, we use the Akaike’s information criterion (AIC; Akaike, 1974) provided by the “glmer” function.

To understand the quality of the extensions, we will use a cross-validation procedure. By the quality of the extensions, we mean the accuracy of predicting correctness of students’ responses in summative assessments using the extensions that is an existing procedure (Ekanadham and Karklin, 2015). We use three MOOCs on Coursera with 1000—4000 active students in each course and three weekly modules from each of these courses. We randomly split students’ responses at summative assessment in each module in three different ways: (a) 50% in the training set and 50% in the test set, (b) 75% in the training set and 25% in the test set, and (c) 95% in the training set and 5% in the test set. Then we fit the extensions and the Rasch model on the training sets and use them to make predictions on the test sets. Then we build a table of confusion with two rows and two columns that report the number of false positives, false negatives, true positives, and true negatives. This table allows us to understand the quality of predictions as well as the improvement of prediction from the Rasch model to the proposed extensions. Using this table, we calculate the accuracy index:

\[
\text{Accuracy} = \frac{TP + TN}{P + N},
\]

Table 1. Overview of the courses used in the cross-validation procedure.

| Course  | Course 2 | Course 3 |
|---------|----------|----------|
| Module 1 |          |          |
| Students | 1609     | 3069     | 4806     |
| Items    | 10       | 10       | 15       |
| Responses| 51550    | 88210    | 141735   |
| Module 2 |          |          |
| Students | 986      | 2050     | 2609     |
| Items    | 10       | 10       | 15       |
| Responses| 30430    | 55100    | 83940    |
| Module 3 |          |          |
| Students | 700      | 1465     | 1735     |
| Items    | 10       | 10       | 15       |
| Responses | 13810  | 36940    | 42750    |

Note: In the table, Course 1 is “Economics for Non-Economists” (Higher School of Economics, n.d.), Course 2 is “Game Theory” (Higher School of Economics, n.d.), and Course 3 is “Introduction to Neuroeconomics: How the Brain Makes Decisions” (Higher School of Economics, n.d.).
where \( TP \) is true positives; \( TN \) is true negatives; \( P \) is all positives; \( N \) is all negatives.

This procedure of splitting the data, fitting the models, making predictions and calculating the accuracy is repeated five more times. Finally, we save all the accuracy indexes and calculate average indexes for each model.

Because the data was coded and anonymized, ethical approval for this research was not required by our institution (KU Leuven).

### 4. Results

We start with the general model presented in Eq. (3). As shown in the first column of Table 2, the estimate of the intercept equals 0.63. To calculate the expected probability that the student gives the correct response to the item, we use the inverse logit function, or antilogit, showing that the expected probability of the correct response is 0.65 (Table 3). The probability of the correct response varies from student to student and especially from item to item. To understand the size of variance among students, we calculated the expected probability of a successful response to an average item, for a student with proficiency of one standard deviation lower, and for a student with proficiency of one standard deviation higher than the average proficiency. These probabilities are 0.46, the antilogit of \((0.63 - 0.78)\), and 0.80, the antilogit of \((0.63 + 0.78)\). The same procedure for items shows us that the probabilities of an average student to give a correct response to an item with difficulty of one standard deviation lower, and to an item with difficulty of one standard deviation higher than the average difficulty are 0.41, the antilogit of \((0.63 - 1.01)\), and 0.84, the antilogit of \((0.63 + 1.01)\), respectively.

| Table 2. Parameters of the extensions. |
|---------------------------------------|
| **Basic Model** | **Extension 1** | **Extension 2** | **Extension 3** | **Extension 4** |
| **Fixed** | | | | |
| Intercept | 0.63 (0.22) ** | 0.27 (0.25) | 0.29 (0.27) | 0.65 (0.27) * |
| Attempt | 0.92 (0.03) *** | 0.90 (0.06) *** | 1.34 (0.07) *** | 1.34 (0.07) *** |
| Class | | | -0.48 (0.03) *** | -0.42 (0.03) *** |
| Attempt * Class | | | -0.33 (0.03) *** | -0.33 (0.03) *** |
| Lect. Act. | | | | |
| Prac. Perf. | | | | |

| **Var** | **SD** | **Var** | **SD** | **Var** | **SD** | **Var** | **SD** |
|---------|--------|---------|--------|---------|--------|---------|--------|
| Random | | | | | | | |
| Intercept | Stud. | 0.61 | 0.78 | 0.84 | 0.92 | 0.92 | 0.96 | 0.74 | 0.86 | 0.59 | 0.77 |
| Item | 1.02 | 1.01 | 1.25 | 1.12 | 1.49 | 1.22 | 1.49 | 1.22 | 1.49 | 1.22 |
| Attempt | Stud. | 0.36 | 0.60 | 0.60 | 0.35 | 0.59 | 0.26 | 0.52 | 0.26 | 0.52 |
| Item | 0.04 | 0.21 | 0.21 | 0.04 | 0.21 | 0.21 | 0.04 | 0.21 | 0.21 | 0.21 |
| AIC | 58779 | 55401 | 54929 | 54487 | 54260 |

Note: ***: \( p < .001; \) **: \( p < .01; \) *: \( p < .05.\)
Probability of the correct response grows with a new attempt to solve the item. To illustrate it, we continue with the first extension presented in Eq. (4). The antilogits derived from Table 2 are summarized in Table 3. The expected probability that the student gives the correct response to the item from the first attempt is .57, while the probabilities that the student gives the correct response to the item at the second and the third attempt are .77 and .89 respectively. However, the effect of attempt varies among students. Thus, the chances for the correct response for a student with average proficiency and an effect of attempt of one standard deviation lower than the average effect of attempt are .57, .64, and .71, at the first, the second, and the third attempt respectively. While the chances to the correct response for a student with average proficiency and an effect of attempt of one standard deviation higher than the average effect of attempt are .57, .86, and .96, at the first, the second, and the third attempt respectively. It is important to note that in the first extension, the standard deviation between items and students is higher than in the general model, 0.92 versus 0.78 for students, and 1.12 versus 1.01 for items (see Table 2).

The effect of attempt also varies from item to item. As presented in Table 3, the chances to the correct response for a student with the average proficiency and the average effect of attempt on an item with an effect of attempt of one standard deviation lower than the average effect of attempt are .57, .73, and .84, at the first, the second, and the third attempt respectively. At the same time, the chances for the correct response for a student with average proficiency and the average effect of attempt on an item with an effect of attempt of one standard deviation higher than the average effect of attempt are .57, .86, and .96, at the first, the second, and the third attempt respectively. The probabilities of correct response on an item with average difficulty.

### Table 3. Probabilities of correct response on an item with average difficulty.

| Effect of Attempt | Maximum Number of Used Attempts | Watched All Video Lectures and Was Productive with Formative Assessments | Attempt |
|-------------------|--------------------------------|-------------------------------------------------|---------|
|                   | Student-Specific               | Item-Specific                                   | 1       | 2       | 3       |
| Basic Model       |                                |                                                 | .65     |         |         |
| Extension 1       | Average                        | Average                                         | .57     | .77     | .89     |
|                   | +1 SD                          |                                                 | .57     | .86     | .96     |
|                   | −1 SD                          |                                                 | .57     | .64     | .71     |
| Extension 2       | Average                        | Average                                         | .57     | .77     | .89     |
|                   | +1 SD                          |                                                 | .57     | .80     | .92     |
|                   | −1 SD                          |                                                 | .57     | .73     | .84     |
|                   | Average                        | Average                                         | .57     | .86     | .96     |
|                   | +1SD                           |                                                 | .57     | .88     | .98     |
|                   | −1SD                           |                                                 | .57     | .83     | .95     |
| Extension 3       | Average                        | Average                                         | .57     | .65     | .71     |
|                   | +1 SD                          |                                                 | .57     | .69     | .79     |
|                   | −1 SD                          |                                                 | .57     | .60     | .62     |
| Extension 4       | Average                        | Average                                         | ≤2      | .66     | .88     | .97     |
|                   |                                |                                                 | >6      | .31     | .39     | .48     |
|                   | Average                        | Average                                         | Yes     | .74     | .92     | .98     |
|                   |                                |                                                 | No      | .32     | .64     | .87     |

Note: In the table, for the basic model and extensions 1 and 2 the student’s proficiency is considered as average.

Probability of the correct response grows with a new attempt to solve the item. To illustrate it, we continue with the first extension presented in Eq. (4). The antilogits derived from Table 2 are summarized in Table 3. The expected probability that the student gives the correct response to the item from the first attempt is .57, while the probabilities that the student gives the correct response to the item at the second and the third attempt are .77 and .89 respectively. However, the effect of attempt varies among students. Thus, the chances for the correct response for a student with average proficiency and an effect of attempt of one standard deviation lower than the average effect of attempt are .57, .64, and .71, at the first, the second, and the third attempt respectively. While the chances to the correct response for a student with average proficiency and an effect of attempt of one standard deviation higher than the average effect of attempt are .57, .86, and .96, at the first, the second, and the third attempt respectively. It is important to note that in the first extension, the standard deviation between items and students is higher than in the general model, 0.92 versus 0.78 for students, and 1.12 versus 1.01 for items (see Table 2).

The effect of attempt also varies from item to item. As presented in Table 3, the chances to the correct response for a student with the average proficiency and the average effect of attempt on an item with an effect of attempt of one standard deviation lower than the average effect of attempt are .57, .73, and .84, at the first, the second, and the third attempt respectively. At the same time, the chances for the correct response for a student with average proficiency and the average effect of attempt on an item with an effect of attempt of one standard deviation higher than the average effect of attempt are .57, .86, and .96, at the first, the second, and the third attempt respectively. It is important to note that in the first extension, the standard deviation between items and students is higher than in the general model, 0.92 versus 0.78 for students, and 1.12 versus 1.01 for items (see Table 2).
attempt are .57, .80, and .92, from the first, the second, and the third attempt respectively. Compared to the basic model, also for the second extension the standard deviation of the probability of the correct response for items and students is higher than in the general model, 0.96 versus 0.78 for students, and 1.22 versus 1.01 for items (see Table 2).

The analysis of the data using Extension 3 (Eq. (6)) teaches us that, as expected, the effect per additional attempt is lower for students who use a higher number of attempts. Let us consider the example in which we calculate the expected probability of the correct response from the first and from the second attempt to an item with the average difficulty for two students. The first student (at least once) used more than six attempts in the summative assessment, while the maximum number of attempts for the second student was not more than two. From the first attempt, the expected probability for the first student is .31, while for the second student it is .66. Probability of the correct response at the second attempt for the same students is .39, for the first student, and .88, for the second student.

Students, active with video lectures and productive with formative assessments, have higher chances to solve items correctly. In the fourth extension (Eq. (7)), presented in Table 2, we see that the effects of video lectures and formative assessments are significant and positive. For instance, the expected probability of the correct response from the first attempt for the student who did not watch video lectures during the week and who was not productive with formative assessments is .32. At the same time, the expected probability of the correct response from the first attempt for the student who watched all video lectures during the week and was productive with formative assessments is .74. It is important to note that including activity with video lectures and productivity with formative assessments to the model explains 20% of the variance among students (in comparison to the third extension).

Finally, we can see that model fit is improving with each further extension, with the AIC decreasing from 58779 for the general model to 54260 for the fourth extension respectively.

To show the improvements in the performance better, we started the cross-validation study from the Rasch model fitted on the first attempts only. The overall accuracy of the predictions using the Rasch model is .710 (Table 4). The basic model, a reformulation of the Rasch model fitted on all responses, has higher accuracy (.739). Including the effects of attempts varying among students improves the accuracy by .029—.768. Then, including the effect of attempts varying among items, effect of class of students, and the explanatory predictors slightly improve the accuracy by .002, .000, and .001 respectively. We can see that the highest improvement of accuracy was from .638 to .738 in Course 2, the lowest improvements were in Course 3, changing from .797 to .821.
5. Discussion & conclusion

In this study several extensions for the Rasch model for applications on MOOC data were presented. The importance of our findings could be described by three points: novelty, relevance for practice, and development of IRT. Firstly, application of IRT in MOOC research is a new field. For instance, the Science Direct search by “title, abstract, and keywords” showed no papers for the combination “massive open online course” or “MOOC” and “psychometrics”, “item response theory”, “IRT”, or “Rasch model”. We can link it to the newness of online learning as a phenomenon in education. Secondly, as can be seen, these extensions allow to measure student’s proficiency in MOOCs more accurately in comparison to the original model. This effect is mainly achieved by involving attempts as a predictor to the proficiency computation process. These findings support the ideas of the explanatory IRT movement started by De Boeck and Wilson (2004) and collaborators, showing that enriching the IRT models with predictors could significantly improve the quality of explanation and understanding of the latent constructs. Finally, the results of this study could be considered as the first steps in transition from the traditional psychometric approaches focused on accurate locating student on the proficiency scale to more flexible approaches for MOOCs, oriented at understanding student’s behavior and linking this behavior to his/her proficiency. We found that the analysis of attempts and student’s individual attempt pattern in MOOCs could be an important source of evidence and might support the analysis of correctness of student’s responses, realized by the traditional psychometric approaches. Linking student’s activity with video-lectures and productivity in formative assessments to correctness of his/her responses in summative assessments allowed us to consider and understand student’s proficiency from the position of the whole module instead of the position of 10 items in summative assessment. We believe it means an important contribution to the field of psychometrics of MOOCs (although the proposed extensions might be applicable in other contexts, it requires additional feasibility studies).

Table 4. Accuracy in predicting correctness.

|                     | Overall M | Overall SD | Course 1 M | Course 1 SD | Course 2 M | Course 2 SD | Course 3 M | Course 3 SD |
|---------------------|-----------|------------|------------|-------------|------------|-------------|------------|-------------|
| Rasch (1st att.)    | .710      | .07        | .695       | .03         | .638       | .02         | .797       | .02         |
| Basic Model         | .739      | .06        | .719       | .03         | .688       | .01         | .811       | .02         |
| Extension 1         | .768      | .04        | .751       | .02         | .734       | .02         | .819       | .02         |
| Extension 2         | .770      | .04        | .753       | .02         | .737       | .02         | .820       | .02         |
| Extension 3         | .770      | .04        | .753       | .02         | .737       | .02         | .821       | .02         |
| Extension 4         | .771      | .04        | .754       | .02         | .738       | .02         | .821       | .02         |

Note: In the table, Course 1 is “Economics for Non-Economists” (Higher School of Economics, n.d.), Course 2 is “Game Theory” (Higher School of Economics, n.d.), and Course 3 is “Introduction to Neuroeconomics: How the Brain Makes Decisions” (Higher School of Economics, n.d.).
The proposed extensions work with test-based assessments, where, according to the theory, the performance is a function of two parameters, student’s proficiency and difficulty of test items. However, assessments in MOOCs could be realized as peer-reviewed assignments where students write an essay or make a design project and then check works of their peers using rubrics with criteria, proposed by the course author. In peer-reviewed assignments, the performance is a function of at least three parameters: student’s proficiency, assignment difficulty and reviewer’s leniency or severity. This example shows us the limitation of our extensions, but the limitation could be considered as a direction for future research. As a solution, we propose to extend the existing psychometric models for peer-reviews such as Many-Facet Rasch Model (Linacre, 1992). The resulting set of extensions will greatly improve measurements in MOOCs.

**Declarations**

**Author contribution statement**

Dmitry Abbakumov: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Piet Desmet: Conceived and designed the experiments.

Wim Van den Noortgate: Conceived and designed the experiments; Wrote the paper.

**Funding statement**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Competing interest statement**

The authors declare no conflict of interest.

**Additional information**

No additional information is available for this paper.

**References**

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6), 716–723.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67 (1), 1–48.

Borsboom, D., 2005. Measuring the Mind: Conceptual Issues in Contemporary Psychometrics. Cambridge University Press, Cambridge.
De Boeck, P., Wilson, M., 2004. Explanatory Item Response Models: a Generalized Linear and Nonlinear Approach. Springer, New York, NY.

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. J. Stat. Software 39 (12), 1–28.

Ekanadham, C., Karklin, Y., 2015. T-SKIRT: online estimation of student proficiency in an adaptive learning system. In: 31st International Conference on Machine Learning. Lille.

Hambleton, R.K., Jones, R.W., 1993. Comparison of classical test theory and item response theory and their applications to test development. Educ. Meas. 12 (3), 38–47.

Higher School of Economics. (n.d.). Economics for Non-economists. Retrieved from Coursera: https://www.coursera.org/learn/ekonomika-dlya-neekonomistov.

Higher School of Economics. (n.d.). Game Theory. Retrieved from Coursera: https://www.coursera.org/learn/game-theory.

Higher School of Economics. (n.d.). Introduction to Neuroeconomics: How the Brain Makes Decisions. Retrieved from Coursera: https://www.coursera.org/learn/neuroeconomics.

Kruyen, P.M., Emons, W.H., Sijtsma, K., 2012. Test length and decision quality in personnel selection: when is short too short? Int. J. Test. 12, 321–344.

Linacre, J.M., 1992. Many-facet Rasch Measurement. MESA Press, Chicago, IL.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison Wesley Publishing Company, Reading, MA.

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. Danish Institute for Educational Research, Copenhagen.

Shah, D., 2017. Coursera’s 2017: Year in Review. Retrieved from Class Central: https://www.class-central.com/report/coursera-2017-year-review/.

Shah, D., 2018. A Product at Every price: a Review of MOOC Stats and Trends in 2017. Retrieved from Class Central: https://www.class-central.com/report/moocs-stats-and-trends-2017/.

Van den Noortgate, W., De Boeck, P., Meulders, M., 2003. Cross-classification multilevel logistic models in psychometrics. J. Educ. Behav. Stat. 28, 369–386.