International Journal of Distance Education Technologies
Volume 20 • Issue 1

An Item Response Theory Approach to Enhance Peer Assessment Effectiveness in Massive Open Online Courses

Minoru Nakayama, Tokyo Institute of Technology, Japan
https://orcid.org/0000-0001-5563-6901

Filippo Sciarrone, Universitas Mercatorum, Italy

Marco Temperini, Sapienza University of Rome, Italy*
https://orcid.org/0000-0002-8597-4634

Masaki Uto, University of Electro-Communications, Japan
https://orcid.org/0000-0002-9330-5158

ABSTRACT

Massive open on-line courses (MOOCs) are effective and flexible resources to educate, train, and empower populations. Peer assessment (PA) provides a powerful pedagogical strategy to support educational activities and foster learners' success, also where a huge number of learners is involved. Item response theory (IRT) can model students' features, such as the skill to accomplish a task, and the capability to mark tasks. In this paper the authors investigate the applicability of IRT models to PA, in the learning environments of MOOCs. The main goal is to evaluate the relationships between some students' IRT parameters (ability, strictness) and some PA parameters (number of graders per task, and rating scale). The authors use a data-set simulating a large class (1,000 peers), built by a Gaussian distribution of the students' skill, to accomplish a task. The IRT analysis of the PA data allow to say that the best estimate for peers' ability is when 15 raters per task are used, with a [1,10] rating scale.

KEYWORDS

Grading Scale, Gaussian distribution, Item Response Theory, Latent Ability, Peer Assessment, Pearson Correlation, Rating Peers, Simulation, Strictness

INTRODUCTION

E-learning technologies have been evolving and expanding at high rates, so Massive Open Online Courses (MOOCs) and Open Educational Resources (OERs) are being rapidly integrated into educational processes by organizations and institutions around the world (West-Pavlov, 2018). The Internet, and in general, digital access to information, has been recognized as the main tool supporting development (Gillwald et al., 2019): that is why several efforts are made, also by international agencies, to promote the availability of network connections in developing countries (Siles, 2020). Suppose the availability and spread of numerous education/training opportunities could help surmount the

DOI: 10.4018/IJDET.313639
*Corresponding Author

abovementioned barrier. In that case, Technology Enhanced Learning (TEL), and Networked Learning in particular, reveal themselves as significant for development beyond their intrinsic educational and pedagogical advantages. Networked education is gaining further confirmation in the current turn of time, while an infectious disease has been spreading and keeps reappearing, forcing long-lasting modifications in the protocols of the worldwide educational systems. Students, and people in general, have to minimize contacts (in presence), and of course, teaching and learning activities have to go on. MOOCs and Open Education are effective and flexible resources to educate, train, and empower populations previously denied actual access to education or online education. MOOCs have been developing for several years to provide learning content for a huge worldwide audience (de Freitas et al., 2015). However, it is still quite frequent that one can enroll in a MOOC and attend the lectures free of charge. On the other hand, easy and inexpensive enrollment is one of the factors significantly producing MOOC dropout rates, together with difficulties in maintaining motivation and engagement.

With respect to motivation and engagement, MOOCs have the same problems as other typologies of distance learning, just made more severe by the extended number of learners. A strong didactic strategy involving the extensive use of assessment, particularly *formative assessment* (Bloom et al., 1971), can be (part of) a solution to such problems. In contrast, the assessment option for MOOCs is known to be limited (Admiraal et al., 2015).

In particular, peer assessment (PA) is available as a powerful strategy to support educational activities and foster learners' success (also when a huge number of learners is implied) (Alcarria et al., 2018). By PA, learners can be exposed to different cognitive experiences: on the one hand, they are requested to perform a task (e.g., answering an open-ended question); on the other hand, they are requested to assess other learners' works, being then involved in cognitive activities of a higher level than just answering (Bloom, 1956).

Moreover, a significant aspect of PA is that it can be delivered at a distance. PA has been already introduced in MOOCs as a learning strategy (Sun et al., 2015). However, the reliability of the assessments, and in general, the applicability of this strategy, are still under discussion by scholars (Alcarria et al., 2018). In practice, there is a question of reliability about the final grade computed by the PA framework if it is based exclusively on peers' marking work. Furthermore, it concerns the reliability of students' grading ability, which is clearly, if not completely, dependent on peers' proficiency in the subject matter of the task to grade. So, studying new PA models suitable to be applied to a MOOC and enhancing the MOOC learning setting is a worthwhile research activity. In particular, one of the directions is how to have a PA system able first to manage, in a computationally feasible way, the big amount of data coming from a PA session in a MOOC. Second, to use such data to maintain reliable student models to use them for automated grading of the tasks/artifacts produced by the learners. Having the above in mind and aiming to study the best configuration of a PA system operating in a MOOC, in this paper, we consider the integration in the PA model of the well-established formal methods described in the item response theory (IRT).

IRT (Baker & Kim, 2004; Lord, 1980) can address the reliability of peer grading using several mathematical models, which provide an Item Response Function (IRF) expressing the probability of a correct response to a given test item. IRF is computed according to the participant's estimated *Proficiency* and the item's characteristics, such as *Difficulty* and *Discrimination*. Recently, IRT models that consider characteristics of raters, such as *Strictness* and *Consistency*, and the item characteristics, have also been proposed and applied to PA.

In general, and in particular, when a large number of students (peers) are involved, the use of PA could be overwhelmingly challenging if certain parameters of the PA were not carefully chosen. Among such parameters, are the *number of raters* (the number of peers requested to mark each peer's task) and the *rating scale* used for the marks. These parameters could represent important factors for the formative effect of PA on the individual student and the accuracy of the assessments given by the peer.

The higher the number of raters assigned to the same task, the higher the assessing work for each peer. In practice, such work could be between *very light* (being less formative) and *very heavy*

(being hardly bearable and, again, less formative). Also, using different rating scales could involve each peer in different levels of complexity during the marking work. For instance, giving a mark between 1 and 10 could be less demanding than giving a mark between 1 and 100.

So, in this paper, we investigate the applicability of some IRT models to PA in a MOOC setting. In particular, we investigate how two IRT parameters, the student's *Ability* and *Strictness*, can be related to two significant aspects of a PA process (and we call them "parameters" as well), the *number of graders* (ng) assigned to peer-evaluate each individual peer-task and the *rating scales* (rss) used to express the evaluation on a task. In short, *Ability* is the student's capability to perform well the assigned task (e.g., to answer correctly to an item), while *Strictness* is the peer's tendency to give consistently lower ratings than the correct ones to the peer-evaluated tasks. In particular, we investigate the following research questions (RQs):

RQ1: Does a correlation exist, and to what extent, between the IRT variable *Ability* and the two parameters ngs and rss?
RQ2: Does a correlation exist, and to what extent, between the IRT variable *Strictness* and the two parameters ngs and rss?
RQ3: Is it possible, via IRT, to estimate the best settings for the abovementioned PA parameters in order to enhance the effectiveness of PA in a MOOC environment?

The paper is structured as follows. The next section shows related works about IRT, MOOCs, and PA. Section 3 (Method) illustrates the method used to combine IRT, PA, and MOOCs in a unique experimental setting. Section 4 (Results) presents the experimental results, and Section 5 (Discussion) discusses the experimental results and some limitations of the study are listed and discussed. Finally, the last section (Conclusions) draws some final remarks and conclusions.

BACKGROUND

PA (Kane & Lawler, 1978; Sadler & Good, 2006) is a well-established learning, assessing, and self-assessing method. It is used in several educational settings and supports the development of knowledge (Li et al., 2010) and meta-cognitive skills in learners (Metcalfe & Shimamura, 1994; Anderson & Krathwhol, 2000). For instance, Tenorio et al. (2016) proposed an extended literature survey concluded that using PA helps improve learners' proficiency in a subject matter.

PA has also been studied in a MOOC environment. Admiraal et al. (2015) proposed an experiment about using self-assessment and peer-assessment in two MOOCs, concluding that PA (and, to some extent, self-assessment) can be an effective tool to implement formative assessment practices in a large-scale environment.

Formanek et al. (2017) report on some experiences involving about 1,000 participants, concluding that the amount of personal participation in the PA protocol is a good predictor for the individual learner's capability and success in the course. In contrast, the grades given to peers by other peers did show a problem of reliability.

Also, Meek et al. (2017) show similar results while describing an experiment of the application of PA in a medical education setting, also providing some interesting further data about the mixed welcome offered to PA by students.

However, to our knowledge, few works address the application of IRT to the PA didactic strategy in a MOOC context. On the one hand, some studies deal with the use of IRT to estimate the grading capability of peers in PA. On the other hand, there are some proposals related to using IRT to model students in MOOCs. We think that our proposal can be significant in bridging the gap between such separated research niches.

Giora et al. (2016) use IRT to identify cheaters in MOOCs. They address the academic dishonesty problem. They developed a general detection method not tailored to a specific form of cheating but

rather based on measuring some aspects of behavior that could be associated with or affected by cheating. The aspects taken into consideration include the amount of interaction with the course resources, the time to answer, the student's ability to answer, two person-fit parameters obtained from IRT-Guttman error, and the standard error of ability estimates.

Uto & Ueno (2016) address the problem of grading reliability in a PA environment, which mainly depends on the rater's characteristics. For this reason, some IRT models that incorporate rater parameters have been proposed to improve grading reliability by accurately estimating some IRT parameters. A hierarchical Bayes model is used for the proposed model to learn some hyperparameters from data. We show the effectiveness of their approach through a set of experiments with real and simulated data.

Uto et al. (2020) propose a new group formation method to maximize PA accuracy using IRT and integer programming. This study proposes an external rater assignment method that assigns a few optimal outside-group raters to each learner after forming groups. We demonstrate that the proposed external rater assignment can substantially improve PA accuracy.

Sterbini & Temperini (2013) presented a system supporting the application of PA able to trace the evolution of students' models during a PA session and to infer grades for the student's answers/tasks. A Bayesian network stores the information from the PA session (grades given, grades taken, and available teacher's grades). A student model is used to infer the grade the corresponding student should receive for her/his task and weigh the grades the student gave other peers during a PA session. Such a system can collect some teachers' grades for a subset of the students' answers/tasks and infer the rest based on the (direct and indirect) influence such teachers' grades have on the whole PA learning process.

De Marsico et al. (2017) present a similar framework based on a different machine learning approach, that is, *K-nearest neighbor*.

Another aspect of interest in the study we are currently presenting is related to certain PA parameters on which the performance of a PA process may depend. The main parameter is usually the *number of peers* assigned/expected to (peer-)evaluate a given task and the *composition of the grading scale* (i.e., the different marks system used by the peers to grade a task).

Regarding the former parameter (number of peers), Cho & MacArthur (2010) conclude that more numerous feedback can help a learner reason about the weaknesses of their artifacts from different viewpoints and to produce a more complex revision. However, to our knowledge, no study has determined the best number of peer assessments that should be requested during PA in a MOOC environment.

Regarding the second parameter, different scale compositions have often been proposed concerning the scoring criteria (instructions for peers about assigning a score to an artifact) in the PA system. In general, using many scoring criteria and scales that are too extended are considered detrimental for the peers' scoring performance. These can be confusing and cause peers to use only a part (usually the lowest and the highest) of the scoring intervals (Miller, 2003).

Nakayama et al. (2020a) examined the feasibility of estimating individual performance for a simulated data set representing a MOOC environment where 1,000 students are supposed to perform a PA session where each peer assesses three other peers' work. For each student, the modeling traits *Ability*, *Consistency*, and *Strictness* are evaluated using the Generalized Many-Facet Rasch Model (GMFRM) (Uto & Ueno, 2018, 2020), and the validity of such calculation is confirmed. This work did not take into account any analysis of the previously mentioned PA parameters (number of graders and grading scales) and provides evidence of the possibility to predict learning performance in the large-scale learning conditions of a MOOC.

Nakayama et al. (2020b) apply IRT to PA in MOOCs with the main goal of estimating the ability of learners by using IRT. In this work, a MOOC composed of 1,000 students is used. The data set is randomly generated using a Gaussian distribution, while several groups of PA sessions are simulated using the Markov Chain Monte Carlo (MCMC) method. The appropriate number of peers is discussed

using several experiments using different groups of raters, such as 3, 5, 9, 15, 30, and 50, using a 10-point grading scale. In this paper, we are deepening the abovementioned investigations using a larger set of possible grading scales ([1, 3]; [1, 5]; [1, 7]; [1, 10]), where smaller scales are taken into consideration besides the 10-point scale.

Besides the abovementioned initial works on integrating IRT in PA, there are several studies defining statistical models based on the Bayesian framework to identify and correct rater biases in PA environments (Bradley, 2019; Goldin & Ashley, 2011; Mi & Yeung, 2015; Piech et al., 2013). These models have been applied to PA data sets provided by a programming course (Bradley, 2019), a MOOC course in human-computer interaction offered by Stanford University (Piech et al., 2013), and another from a MOOC course in science, technology, and society in China offered by Hong Kong University (Mi & Yeung, 2015). The core idea of these models is to parameterize students' traits, including *Ability*, Consistency, and Strictness, and estimate them using a PA data set. Unlike these approaches, our study focused on the IRT-based approach because IRT is a traditional and sophisticated theory that has long been used in educational and psychometric measurement fields.

OUR STUDY METHOD

In this section, we describe the following:

1. The method we adopted to generate the simulated MOOC and the related PA session. Note that all data analysis was based on simulated data sets generated using the K-OpenAnswer system (Sciarrone & Temperini, 2020). No real experimentation was conducted to gather data.
2. The IRT setting using the GMFRM.
3. The IRT parameters estimation using the GMFRM over the PA samples. The goal was to study how to evaluate the effectiveness of the PA technique used concerning features such as the number of peers requested to grade a given task and the grading scale used.

The Peer Assessment Samples

The web platform K-OpenAnswer carries out What-if analyses (Arsham & Kahn, 1990) on simulated MOOCs based on PA as the didactic strategy. The K-OpenAnswer system can generate several PA-simulated sessions. The K-OpenAnswer system (Sciarrone & Temperini, 2020) was used to produce all data. In order to generate a MOOC together with a PA session, the user has to input the following parameters:

- **The total number N of students belonging to the simulated MOOC:** This number can be very large (up to many thousands) depending on the client computer RAM capability (it is a JavaScript application). In our study, we simulated a MOOC composed of $N = 1,000$ students.
- **The number *GP* of grading peers:** In a PA session, it represents the number of peers assigned to grade a task for the assessment. This number is the same for each task to be assessed.
- **The grading scale *GS*** : This parameter is the range of the grading scale that peers can use to express their assessment in a PA session. For instance, a rating scale is $GS = \left[1,2,\ldots,10\right]$. The GS variable is a continuous and real variable.
- **The statistical parameters:** The K-OpenAnswer user is required to set the following parameters to characterize the properties they want the statistical distribution data set to be built on:
  - **The type of statistical distribution of the marks assigned by peers can be either Gaussian or uniform:** In the Gaussian case, the user has to enter the mean μ and the variance σ of the distribution.
  - **The peers selection procedure:** In a PA session, once the students have accomplished the tasks, each task is assigned to a set of $m = GP$ peers, who will perform the peer assessment

on that task. The selection of such peers must be conducted systematically and consistently, taking into account each task. In the K-OpenAnswer system, either a circular or a random selection modality can be chosen by the user. In the former modality, the N students are indexed from 1 to N, and for the $i-th$ student's task, the next m students are selected.[1]

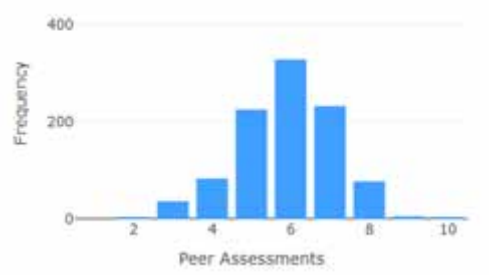
Table 1 shows the Gaussian parameters of all data sets, together with the GP and GS variables. Figure 1 shows the peers' distribution of one of the samples generated by the K-OpenAnswer system. As for the Gaussian data generation algorithm, the Box-Muller transform method was used (Box & Muller, 1958).

As shown in Table 1, 24 data sets were built. Each one comprised of the peers' assessments according to each combination of GS and GP parameters. On these data sets, the GPFRM parameter estimation procedure was applied, as shown in the next section.

Table 1. The statistics of the data sets produced by the K-OpenAnswer System that we used for the experimentation

Data set	Grading scale	# Grading peers	μ	σ^2
1–6	[1, 3]	3, 5, 9, 15, 30, 50	1.5	0.2
7–12	[1, 5]	3, 5, 9, 15, 30, 50	2.5	0.9
13–18	[1, 7]	3, 5, 9, 15, 30, 50	4	1.2
19–24	[1, 10]	3, 5, 9, 15, 30, 50	5.5	1.63

Figure 1. The data set (grades) produced by the K-OpenAnswer system (Note: $N = 1,000$ students, $GP = 5$, and the Gaussian distribution $\mu = 5.5$ and $\sigma = 1.63$)



IRT Setting

The IRT models define the probability of an examinee to correctly respond to a test item as a function of their *Ability* level and of some other item parameters, which vary according to the model taken into consideration.

Consequently, the mark an examinee obtains for a given test item depends on the ability level and the characteristics of the item itself (and on the IRT model used), which determines which parameters of the items are to be included. IRT has traditionally been applied to test items for which responses can be scored as correct or incorrect. However, in recent years, there have been several attempts to apply *many-facet* polytomous IRT models (Eckes & Jin, 2021; Jin & Wang, 2018; Linacre, 1989; Shin et al., 2019; Uto & Ueno, 2018, 2020; Wilson & Hoskens, 2001) to performance assessments, including PA applied to open answer tasks (Chan et al., 2017; Hua & Wind, 2019; Jin & Wang,

2017; Kaliski et al., 2013; Tavakol & Pinner, 2019; Uto & Ueno, 2016; Uto et al., 2020). This is the case of our investigation, where a PA session is based on (simulated) students who accomplish a task based on an open answer assignment. So, some typical rater characteristics of the IRT models to take into account are:

- **Strictness:** The tendency to consistently give low ratings.
- **Consistency:** The extent to which the rater does not grade similar performances differently.
- **Range restriction:** The tendency to overuse a limited number of rating categories. Special cases of range restrictions are: (i) the central tendency, namely, a tendency to overuse the central categories, and (ii) the extreme response tendency, a tendency to prefer endpoints of the response scale (Elliot et al., 2009).

Representative tasks or item characteristics to take into account are:

- **Difficulty:** More difficult tasks tend to receive lower ratings.
- **Discrimination:** The extent to which different level task assessment corresponds to the different outcome quality of the tasks.

The GMFRM has been proposed as one of the latest IRT models that can estimate the peer's abilities while considering the abovementioned rater and item characteristics (Uto & Ueno, 2019, 2020). This model provides the probability P_{ijrk} that a peer's rater r assigns the grade k to participant j's performance for item (or performance task) i, that is:

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\alpha_r \alpha_i \left(\theta_j - \beta_i - \beta_r - \beta_{rm} \right) \right]}{\sum_{l=1}^{k} \exp \sum_{m=1}^{l} \left[\alpha_r \alpha_i \left(\theta_j - \beta_i - \beta_r - \beta_{rm} \right) \right]} \tag{1}$$

where θ_j is the latent ability of participant j, α_r reflects the consistency of rater r, α_i is a discrimination parameter for item i, β_i is the difficulty of item i, β_r is the strictness of rater r, and β_{rm} is a step difficulty parameter denoting the difficulty of transition between scores $m-1$ and m in the rater r. Here, $\prod_i \alpha_i = 1$, $\sum_i \beta_i = 0$, $\beta_{r1} = 0$, and $\sum_{k=2}^{K} \beta_{rk} = 0$ are given for model identification. In the case where the number of items is one, that is our case, α_i and β_i can be ignored because the model identification constraints restrict the values of $\alpha_{i=1} = 1$ and $\beta_{i=1} = 0$, so obtaining for our study:

$$P_{jrk} = \frac{\exp \sum_{m=1}^{k} \left[\alpha_r \left(\theta_j - \beta_r - \beta_{rm} \right) \right]}{\sum_{l=1}^{k} \exp \sum_{m=1}^{l} \left[\alpha_r \left(\theta_j - \beta_r - \beta_{rm} \right) \right]} \tag{2}$$

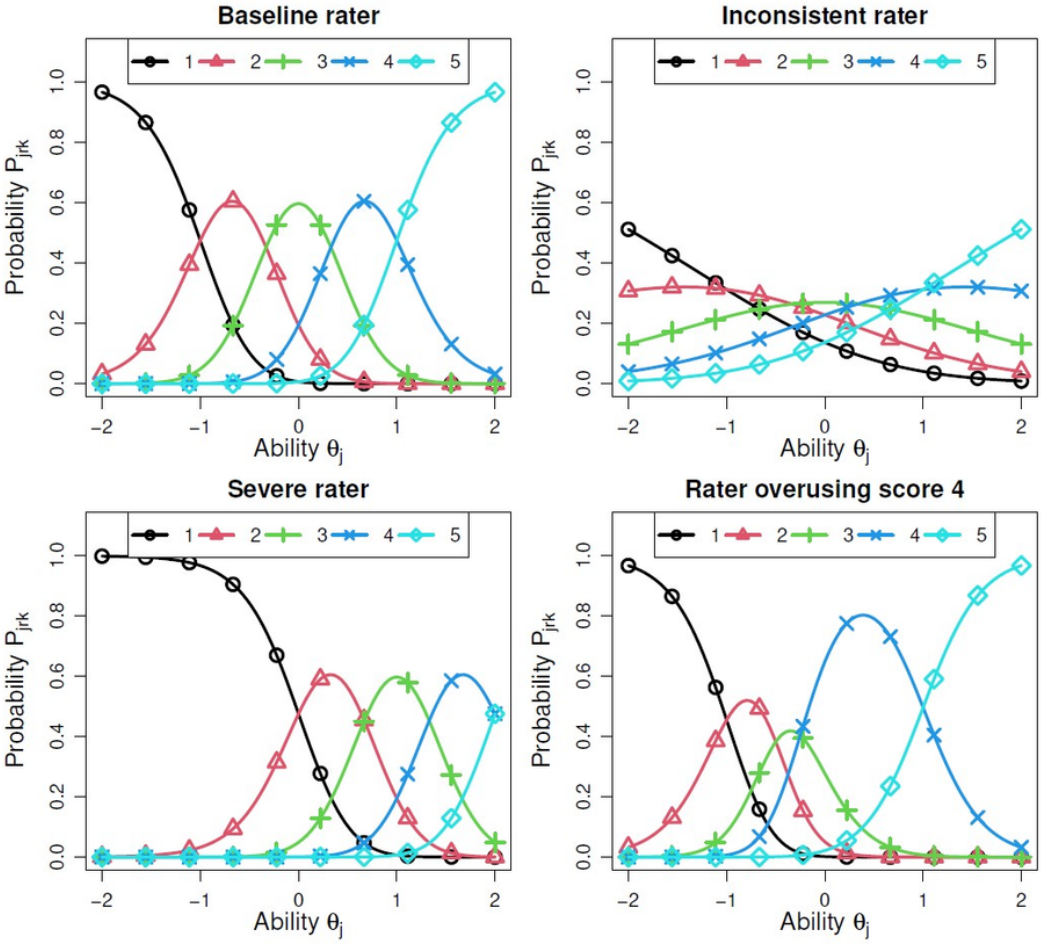
To explain how the GP parameters in the IRT model reflect the three abovementioned typical rater characteristics, Figure 2 shows several curves of the Item Category Response Function (ICRF) of the GMFRM, which are drawn by plotting the probability P_{jrk} concerning four raters having

different parameters. In the figure, the x-axis reports the latent score θ_j, while the y-axis shows the probability P_{jrk}: in summary, participants with higher ability θ_j tend to obtain higher scores.

In the GMFRM, rater consistency is represented by α_r, with lower values indicating smaller differences in response probabilities between score categories. This can be confirmed by comparing the ICRF on the top-left and the top-right in Figure 2, which are drawn using different α_r values. These figures suggest that the marks given by a peer with a lower consistency parameter will be unreliable because the peer tends to assign different scores to participants with similar ability levels.

The GMFRM represents the severity of each rater through the β_r parameter. The ICRF shifts to the right as this parameter increases, indicating that raters with high β_r values tend to consistently assign low scores. In Figure 2, we can confirm that the ICRF on the bottom-right, which is drawn using a high β_r value, shifts to the right overall.

Figure 2. Curves of item category response function (ICRF) for four raters with different rater parameters



Moreover, the GMFRM reflects the range restriction characteristic as β_{rm}. The closer $\beta_{r(m+1)}$ and β_{rm} are, the lower the overall probability of responding with score category m. Conversely, the higher the difference $\beta_{r(m+1)} - \beta_{rm}$ becomes, the higher the response probability for score category m. In Figure 2, the ICRF on the bottom-right has been drawn using some β_{rm} values in which $\beta_{r3} - \beta_{r2}$ and $d_{r4} - d_{r3}$ are small and $\beta_{r5} - \beta_{r4}$ is large. Thus, in the ICRF, response probabilities for scores 2 and 3 decrease, whereas those for score 4 increase, representing a range restriction characteristic with the overuse of score 4, while avoiding scores 2 and 3.

The GMFRM can estimate these rater parameters with ability values θ_j from a collection of PA samples. Thus, the GMFRM helps us to investigate rater characteristics in PA environments and to accurately measure participants' abilities while removing the influence of those effects.

IRT Parameter Estimation

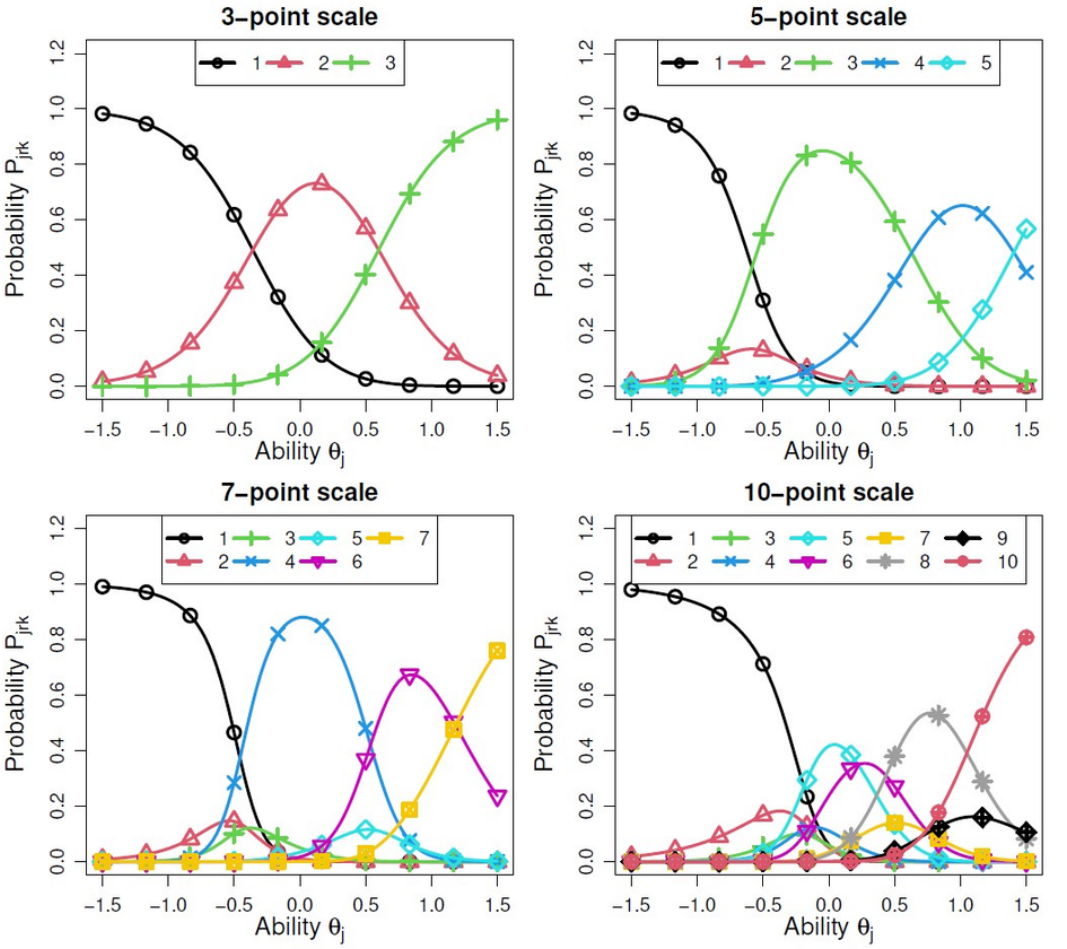
The IRT model parameters are determined as the Expected A Posteriori (EAP) estimation model, where a form of Bayesian estimation, such as the MCMC, is generally used (Fox, 2010; Uto, 2019). In particular, we used the Metropolis-Hastings-within-Gibbs sampling algorithm as the MCMC algorithm for IRT models (Patz & Junker, 1999). Moreover, in recent years, the No-U-Turn (NUT) sampler, an extension of the Hamiltonian Monte Carlo (HMC) that eliminates hand-tuned parameters, has been proposed (Hoffman & Gelman, 2014). Consequently, because the Stan software[2] package implements a NUT-based HMC easy, this algorithm has recently been used for parameter estimations in various statistical models, including IRT models (Jiang & Carter, 2019; Luo & Jiao, 2018). For the aforesaid reasons, we used a NUT-based MCMC algorithm to estimate the model parameters.

We calculated the EAP estimates as the average of the parameter samples obtained from 2,000 to 3,000 periods of three independent MCMC chains. Furthermore, the standard Gaussian distribution was used as the prior distributions. Finally, the MCMC was run using a 2.5 GHz 14 core processor (Intel Xeon W). The convergence was confirmed using the Gelman-Rubin statistics $\hat{R}$ (Gelman & Rubin, 1992), which was less than 1.1 for all parameters, indicating that the MCMC runs converged. The elapsed times for each condition were measured in seconds.

RESULTS

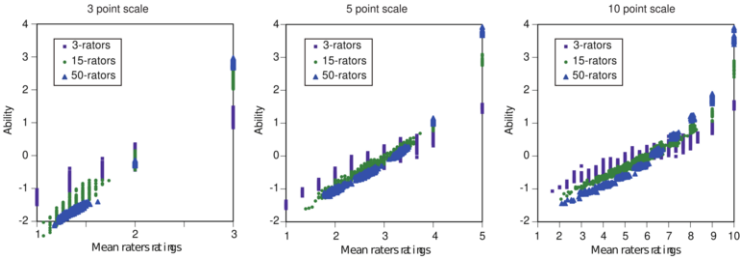
Through the GMFRM, introduced in Section 3.2 (IRT Setting), we estimated the assessing behaviors in the simulated PA session according to the data sets described in Section 3.1 (The Peer Assessment Samples). Figure 3 shows some examples of the curves of the ICRF in the case of five peers assessed by each student using four grading scales (3, 5, 7, 10). In each figure, the horizontal x-axis represents the Latent Ability θ, while the vertical y-axis is the response probability for each rating category. Generally, a good level of discrimination is obtained between the extremes, as in the case of the 10-point scale, where 10=good, and 1=bad, whereas the middle ranges are more overlapped while the grading scale grows in points. Most curves are flattened, and any level of ability cannot rate well the targeted performance. However, most parameters can be estimated well. When many points for rating are given, the rater's marking is inappropriate, as shown in Figure 3. This phenomenon is our motivation to optimize the measure setting as noted in the Introduction. When PA marks were given, the probability of each mark may be broadly distributed and sometimes flattened. If the distributions were strictly controlled, the figure might be irregular.

Figure 3. Curves of item category response function (ICRF) for a number of raters as n = 5 using four types of grading scales ([1, 3]; [1, 5]; [1, 7]; [1, 10]) (Note: The x-axis reports Ability measured in the standard range [–1:5; +1:5], while the y-axis shows the probability)



The relationship between mean rater's ratings and the estimated *Ability* is summarized in the three scatter plots of Figure 4, where we show the data related to 3, 5, and 10-point grading scales. In each image, the comparison is shown to 3, 15, and 50 grading peers.

Figure 4. Relationship between mean rating scores and ability between 3 and 50 raters



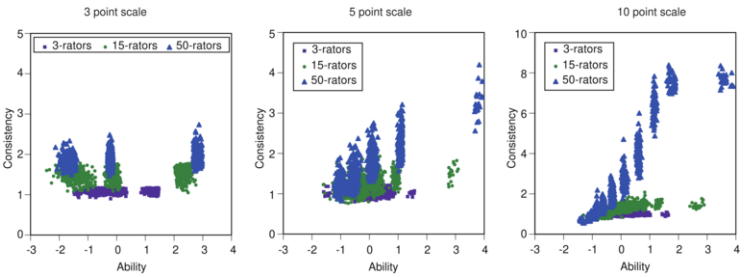
All results show strong positive correlations, and all deviations of mean ratings and *Ability* are reduced, while the number of graded peers is from 3 to 50. So, the distribution of mean rater's ratings is influenced by the number of rated peers. The standard deviation (STD) of means is discrete for three raters, and they are variable values when the number of raters increases.

Table 2. Correlation coefficients with the number of raters using 3- to 10-point scales

# Raters	Ability-Given rating 3 5 7 10	Strictness-Provided rating 3 5 7 10
3	0.99 0.97 0.96 0.95	–0.94 –0.82 –0.75 –0.63
5	0.99 0.97 0.96 0.96	–0.89 –0.78 –0.70 –0.54
9	0.99 0.97 0.97 0.96	–0.85 –0.69 –0.64 –0.56
15	0.99 0.97 0.96 0.96	–0.81 –0.65 –0.65 –0.64
30	0.99 0.94 0.94 0.95	–0.71 –0.67 –0.64 –0.65
50	0.99 0.93 0.92 0.93	–0.62 –0.67 –0.72 –0.73

The correlation coefficients between the mean rater's ratings and the estimated *Ability* are summarized in Table 2. The coefficient may be independent with respect to the number of raters, although the distributions are quite different in Figure 4. When the number of peers increased, the STD of rating scores became small over all the rating values. Changes in the *Consistency* in the same conditions are summarized in Figure 5.

Figure 5. Relationship between ability and consistency between 3 and 50 raters



In Figure 5, the $x-$axis indicates the estimated *Ability*, while the y-axis indicates the estimated *Consistency*. The scales of the y-axis are spread according to the number of points in the grading scale. The range of the *Consistency* value is comparable between the grading scales of 3 and 5; the STD gets larger when a 10-point grading scale is used. Then, a sufficiently high number of scale points seems needed, as the *Ability* is not estimated continuously when a 3-point scale is introduced. When looking at the number of raters in each figure, the deviation of the *Consistency* seems to increase with their number. Therefore, both the grading scale density and the raters number appear to influence the deviation of *Consistency*. In addition, *Consistency* for rating may be influenced by rating steps and rated people.

The opposite index, such as the *Strictness* in rating, is summarized with their rating scores in Figure 6. The $x-$axis indicates the generated rating scores for each individual, while the $y-$axis

indicates the estimated *Strictness* values. When the number of rating peers is three, the STD of rating grades is sparse in this data generation procedure. All of the relationships show a strong negative correlation. As known generally, the deviations decrease with the number of raters, as shown in Figure 4. The correlation coefficients are compared with the number of raters and rating scales in Table 2. The change of coefficients across the number of point scales differs among the number of raters. The relationships between the estimated *Ability* and the *Strictness* are also summarized using the same format in Figure 7. The relationships change with the number of rating scales, as shown in the figure. When a 10-point scale is used, their correlation appears, although the *Strictness* simply deviated for the level of *Ability* using 3- and 5-point scales. So, the above results show that the plausible parameters have been estimated sufficiently, confirming the feasibility of applying PA to an IRT model such as GMFRM.

Figure 6. Relationship between mean rating scores and strictness for 3, 15, and 50 raters

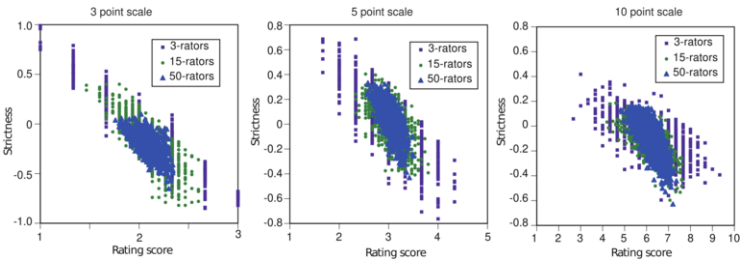
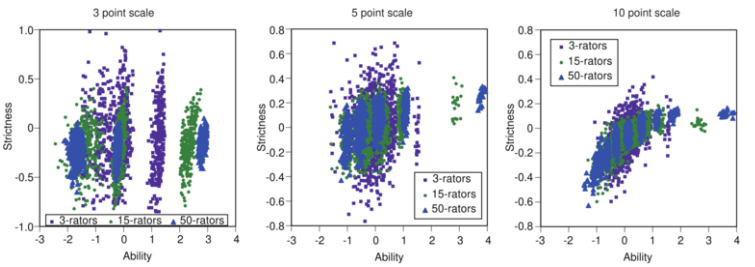


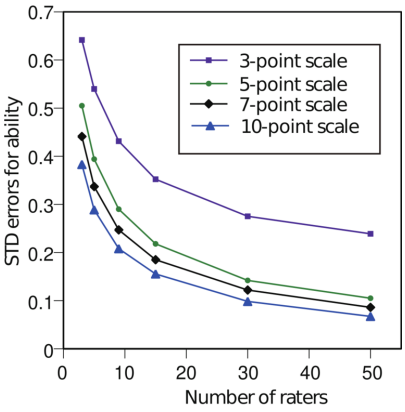
Figure 7. Relationship between ability and strictness for 3, 15, and 50 raters



DISCUSSION

In this study, parameter estimation was conducted for PA in 24 different conditions, where the number of peers assigned to the same evaluation task (number of raters) varied in (3, 5, 9, 15, 30, 50) using rating scales of: 3, 5, 7, and 10-points. The aim was to help tune the number of raters and the grading scale while designing a PA session in a MOOC context. We computed the parameter estimation for *Ability* (θ) along the different conditions, looking for a minimization of the STD error. The progressive changes are summarized in Figure 8.

Figure 8. Standard deviation (STD) errors of ability with the number of raters



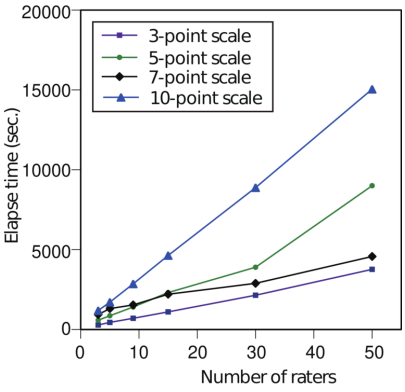
As the figure shows, while the number of raters increases, we note that:

- The curve related to a 3-point scale always remains on a higher STD error.
- The other scales are on a lower level of STD error, with not a particularly severe difference.
- For all the scales, the STD error decreases steadily.

So, as far as the STD error on the estimation of *Ability* is concerned, we could suggest a twofold preliminary conclusion:

1. On the one hand, increasing rater numbers seems to have a distinct effect on the quality of estimation for each of the grading scales. In particular, the estimation ability increases at a higher speed, while the number of peers increases up to 15, and a lower speed for greater values. This would suggest 15 as an interesting branch point.
2. On the other hand, comparing the performances of the grading scales, we see that while the 3-point scale is consistently worse than the others, 5, 7, and 10-point are comparable, with the 10-point scale slightly better.

Figure 9. Elapsed time, in seconds, for the estimation of ability, according to different numbers of raters



We also considered the time elapsed during the *Ability* estimation, summarized in Figure 9. There is a clear dependence of the computational time on the number of raters. It is quite predictable here, that the 3-point scale performs better, while the 10-point scale needs more time. On the other hand, we see a branching point around 15 raters, up to where the 5-point scale is performing slightly better, or equal, against the 7-point scale.

Together with the observations proposed concerning Figure 8, the data in Figure 9 suggest that a comparatively best performance of the PA in our simulated MOOC is obtained using 15 as the number of grading peers per task, with a 5-point grading scale. This conclusion is obtained in the framework of a What-if analysis based on simulated PA data: to some extent, we could conclude that for a MOOC class similar to the one depicted by the simulated data, our conclusions appear to be reasonable.

It is worth noticing, in conclusion, that for different sample (obtained by different Gaussian distributions (cfr. Section 3.1 The Peer Assessment Samples), we might have different suggestions for the mentioned PA parameters. So we have shown a process of analysis, based on an IRT model, which can be applied in different contexts to optimize some aspects of PA in the large-scale classes of a MOOC. On the other hand, such an abstract approach certainly needs further validation by real-world experiments, which will be a subject for our further study.

Limitations of the Study, and Future Work

The study we presented undoubtedly has some limitations that might hinder its effective replication in real-world experiments. In the following, we list some of these limitations that the reader may have already uncovered and point out some plans for future work in relation to them.

The first limit is, of course, in the synthetic nature of the data set we (computed and) used. Before applying to a given real MOOC class, our process of definition of suitable values for the PA parameters, one should first build a PA data set from previous PA sessions and be reasonably sure that such real data model the class. In other words, our process would be reproducible in the real world, only assuming that the (real-world) data set truly represents the peers. Future work should follow two paths: (1) On the one hand, field experimentation should be conducted to confirm the process feasibility and effectiveness in determining the optimal PA parameters; and (2) a comprehensive software system (of which we currently have only detached modules) should support the teacher(s) in the peer-model extraction from previous sessions of PA and in the definition and application of the PA parameters according to the analysis of the existing data set.

The fact that we used a single Gaussian distribution is also to be considered a limitation here. By creating more than one data set, we might have proposed and compared different What-if analyses by different parameter settings for the distribution, which would enrich the study. However, in this paper, we concentrated on defining the general shape of the analysis process, and we considered that more parallel applications of the process would have made the paper (even) more cumbersome, without adding decisive results in exchange.

A third limitation of the study is related to the values and intervals used for the two PA parameters we analyzed. We have discussed their ratio, referring to existing literature. However, there are different choices that we have not considered and might influence the final results of the analysis. For instance, we have not considered very large grading scales, our larger one $[1,10]$. This is partly to ease the production, and the readability, of the investigation and partly because previous studies (Miller, 2003) have pointed out that excessive mark intervals would end up confusing, rather than empowering, the peer. However, other intervals, such as $[0,1]$ (the very basic accept/not accept judgement), or $[1,18]$ and $[1,30]$ (that might be significant in the educational systems of some countries), could be of interest for further work on our process of analysis. Similar observations could be proposed for the number of grading peers: the gap between 15 and 50 is quite large, and further work should consider using at least two more values between them.

A final limitation of our study is that the described process of analysis can only be performed on relatively big numbers, making the application of our process not realistic in smaller classes. A way to overcome the above limits would be to allow for the aggregation of data from several previous sessions of PA held in the same class. However, this aspect is a matter for quite challenging future work, where the concepts of Group Decision-making and completion of the created data set (Alonso et al., 2008; 2009) seem applicable.

CONCLUSION

In this paper, we examined the feasibility and effectiveness of applying an IRT model in a PA framework to determine some optimal values for certain parameters of the PA process in a large-scale (MOOC) educational environment. The study we presented consisted of a What-if analysis based on a MOOC class composed of 1,000 peers. The peers' capability to accomplish a required task (e.g., to answer a question, having it then evaluated by peers) was modeled using a Gaussian distribution.

The results show that it is possible to study, by means of typical IRT variables, the performance of a PA system, for instance, given producing automated grading. In particular, we concluded that for a MOOC class similar to the synthetic sample we computed, a reasonable PA parameter setting is the one assigning 15 to a *number of grading peers per task* and $[1, 5]$ to *grading scale composition*.

As mentioned in the Discussion section, many limitations and considerable issues exist. The detailed discussion and improvement for these issues will be subjects of our further study.

COMPETING INTERESTS

The authors of this publication declare there is no conflict of interest.

FUNDING

This research was partially supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (KAKEN, 20H01718: 2020-2022). This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Admiraal, W., Huisman, B., & Pilli, O. (2015). Assessment in massive open online courses. *The Electric Journal of e-Learning, 13*(4), 207-216.

Alcarria, R., Bordel, B., de Andrès, D. M., & Robles, T. (2018). Enhanced peer assessment in MOOC evaluation through assignment and review analysis. *International Journal of Emerging Technologies in Learning*, *13*(1), 206–219. doi:10.3991/ijet.v13i01.7461

Alonso, S., Chiclana, F., Herrera, F., Herrera-Viedma, E., Alcala-Fdez, J., & Porcel, C. (2008). A consistency-based procedure to estimate missing pairwise preference values. *International Journal of Intelligent Systems*, *23*(1), 155–175. doi:10.1002/int.20262

Alonso, S., Herrera-Viedma, E., Chiclana, F., & Herrera, F. (2009). Individual and social strategies to deal with ignorance situations in multi-person decision making. *International Journal of Information Technology & Decision Making*, *8*(2), 313–333. doi:10.1142/S0219622009003417

Anderson, L., & Krathwohl, D. (2000). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn & Bacon.

Arsham, H., & Kahn, A. (1990). "What-if" analysis in computer simulation models: A comparative survey with some extensions. *Mathematical and Computer Modelling*, *14*, 101–106. doi:10.1016/0895-7177(90)90156-H

Baker, F., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques. (Statistics, textbooks and monographs)*. Marcel Dekker. doi:10.1201/9781482276725

Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook: The cognitive domain*. David McKay.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. McGraw-Hill Inc.

Box, G., & Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, *29*(2), 610–6011. doi:10.1214/aoms/1177706645

Bradley, S. (2019). Addressing bias to improve reliability in peer review of programming coursework. *In Proceedings of 19th Koli Calling International Conference on Computing Education Research*. Association for Computing Machinery. doi:10.1145/3364510.3364523

Chan, S., Bax, S., & Weir, C. (2017). *Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors. Technical Report, IELTS Research Reports*. Online Series.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction*, *20*(4), 383–394. doi:10.1016/j.learninstruc.2009.08.006

de Freitas, S., Morgan, J., & Gibson, D. (2015). Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision. *British Journal of Educational Technology*, *46*(3), 455–471. doi:10.1111/bjet.12268

De Marsico, M., Sciarrone, F., Sterbini, A., & Temperini, M. (2017). Modeling a peer assessment framework by means of a lazy learning approach. In T. C. Huang, R. Lau, Y. M. Huang, M. Spaniol, & C. H. Yuen, (Eds.) *Emerging technologies for education*. In *Proceedings of SETE 2017. Lecture notes in computer science* (vol.10676, pp. 336-345). Springer. doi:10.1007/978-3-319-71084-6_38

Eckes, T., & Jin, K.-Y. (2021). Examining severity and centrality effects in TestDaF writing and speaking assessments: An extended Bayesian many-facet Rasch analysis. *International Journal of Testing*, *21*(3-4), 131–153. doi:10.1080/15305058.2021.1963260

Elliot, M., Haviland, A., Kanouse, D., Hambarsomian, K., & Hays, R. (2009). Adjusting for subgroup differences in extreme response tendency in ratings of health care: Impact on disparity estimates. *Health Services Research*, *44*(2 Pt 1), 542–561. doi:10.1111/j.1475-6773.2008.00922.x PMID:19040424

Formanek, M., Wenger, M. C., Buxner, S. R., Impey, C. D., & Sonam, T. (2017). Insights about large-scale online peer assessment from an analysis of an astronomy MOOC. *Computers & Education*, *113*, 243–262. doi:10.1016/j.compedu.2017.05.019

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. Springer. doi:10.1007/978-1-4419-0742-4

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. doi:10.1214/ss/1177011136

Gillwald, A., Calandro, E., Chair, C., Mothobi, O., & Rademan, B. (2019*). Understanding digital access and use in the global south*. Technical Report 108336-001, Research ICT Africa. https://idl-bnc-idrc.dspacedirect.org/bitstream/handle/10625/58175/58311.pdf

Giora, A., Sunbok, L., Zhongzhou, C., & Pritchard, D. E. (2016). Detecting cheaters in MOOCs using item response theory and learning analytics. In *Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalisation (UMAP 2016)* (vol. 1618, pp. 1-10). Halifax, Canada. CEUR.

Goldin, I. M., & Ashley, K. D. (2011). Peering inside peer review with Bayesian models. In *Proceedings of International Conference on Artificial Intelligence in Education,* (pp. 90-97). Springer-Verlag. doi:10.1007/978-3-642-21869-9_14

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hua, C., & Wind, S. A. (2019). Exploring the psychometric properties of the mind-map scoring rubric. *Behaviormetrika*, *46*(7), 73–99. doi:10.1007/s41237-018-0062-z

Jiang, Z., & Carter, R. (2019). Using Hamiltonian Monte Carlo to estimate the loglinear cognitive diagnosis model via Stan. *Behavior Research Methods*, *51*(2), 651–662. doi:10.3758/s13428-018-1069-9 PMID:29949073

Jin, K. Y., & Wang, W. C. (2017). Assessment of differential rater functioning in latent classes with new mixture facets models. *Multivariate Behavioral Research*, *52*(3), 391–402. doi:10.1080/00273171.2017.1299615 PMID:28328280

Jin, K. Y., & Wang, W. C. (2018). A new facets model for rater's centrality/extremity response style. *Journal of Educational Measurement*, *55*(4), 543–563. doi:10.1111/jedm.12191

Kaliski, P. K., Wind, S. A., Engelhard, G. Jr, Morgan, D. L., Plake, B. S., & Reshetar, R. A. (2013). Using the many-faceted Rasch model to evaluate standard setting judgments. *Educational and Psychological Measurement*, *73*(3), 386–411. doi:10.1177/0013164412468448

Kane, L., & Lawler, E. (1978). Methods of peer assessment. *Psychological Bulletin*, *85*(3), 555–586. doi:10.1037/0033-2909.85.3.555

Li, L., Liu, X., & Steckelberg, A. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology*, *41*(3), 525–536. doi:10.1111/j.1467-8535.2009.00968.x

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. MESA Press.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Erlbaum Associates.

Luo, Y., & Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, *78*(3), 384–408. doi:10.1177/0013164417693666 PMID:30140099

Meek, S. E., Blakemore, L., & Marks, L. (2017). Is peer review an appropriate form of assessment in a MOOC? Student participation and performance in formative peer review. *Assessment & Evaluation in Higher Education*, *42*(6), 1000–1013. doi:10.1080/02602938.2016.1221052

Metcalfe, J., & Shimamura, A. (1994). *Metacognition: knowing about knowing*. MIT Press. doi:10.7551/mitpress/4561.001.0001

Mi, F., & Yeung, D.-Y. (2015). Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 454-460). AAAI Press. doi:10.1609/aaai.v29i1.9210

Miller, P. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment & Evaluation in Higher Education*, *28*(4), 383–394. doi:10.1080/0260293032000066218

Nakayama, M., Sciarrone, F., Uto, M., & Temperini, M. (2020a). Estimating student's performance based on item response theory in a MOOC environment with peer assessment. In *Proceedings of 10th International Conference in Methodologies and Intelligent Systems for Technology Enhanced Learning* (pp. 1-10). Springer.

Nakayama, M., Sciarrone, F., Uto, M., & Temperini, M. (2020b). Impact of the number of peers on a mutual assessment as learner's performance in a simulated MOOC environment using the IRT model. In *Proceedings of 24th International Conference Information Visualisation (IV)* (pp. 483-487). IEEE. doi:10.1109/IV51561.2020.00084

Patz, R. J., & Junker, B. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342–366. doi:10.3102/10769986024004342

Piech, C., Huang, J., Chen, Z., Do, C. B., Ng, A. Y., & Koller, D. (2013). Tuned models of peer assessment in MOOCs. In *Proceedings of International Conference of the Educational Data Mining Society* (pp.153-160). https://files.eric.ed.gov/fulltext/ED599096.pdf

Sadler, P., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, *11*(1), 1–31. doi:10.1207/s15326977ea1101_1

Sciarrone, F., & Temperini, M. (2020). K-OpenAnswer: A simulation environment to analyze the dynamics of massive open online courses in smart cities. *Soft Computing*, *24*(15), 11121–11134. doi:10.1007/s00500-020-04696-z

Shin, H. J., Rabe-Hesketh, S., & Wilson, M. (2019). Trifactor models for multiple-ratings data. *Multivariate Behavioral Research*, *54*(3), 360–381. doi:10.1080/00273171.2018.1530091 PMID:30919664

Siles, I. (2020). *A transnational history of the internet in Central America, 1985–2000*. Palgrave Macmillan. doi:10.1007/978-3-030-48947-2

Sterbini, A., & Temperini, M. (2013). Analysis of open answers via mediated peer-assessment. In *Proceedings of 17th International Conference on System Theory, Control and Computing* (pp. 663-668). ICSTCC. doi:10.1109/ICSTCC.2013.6689036

Sun, D. L., Harris, N., Walther, G., & Baiocchi, M. (2015). Peer assessment enhances student learning: The results of a matched randomized crossover experiment in a college statistics class. *PLoS One*, *10*(12), e0143177. doi:10.1371/journal.pone.0143177 PMID:26683053

Tavakol, M., & Pinner, G. (2019). Using the many-facet Rasch model to analyse and evaluate the quality of objective structured clinical examination: A non-experimental cross-sectional design. *BMJ Open*, *9*(9), 1–9. doi:10.1136/bmjopen-2019-029208 PMID:31494607

Tenorio, T., Bittencourt, I., Isotani, S., & Silva, A. (2016). Does peer assessment in online learning environments work? A systematic review of the literature. *Computers in Human Behavior*, *64*, 94–107. doi:10.1016/j.chb.2016.06.020

Uto, M. (2019). Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), Lecture notes in computer science, Vol. 11625. *Artificial Intelligence in Education. AIED 2019*. Springer. doi:10.1007/978-3-030-23204-7_41

Uto, M., Nguyen, D. T., & Ueno, M. (2020). Group optimization to maximize peer assessment accuracy using item response theory and integer programming. *IEEE Transactions on Learning Technologies*, *13*(1), 91–106. doi:10.1109/TLT.2019.2896966

Uto, M., & Ueno, M. (2016). Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, *9*(2), 157–170. doi:10.1109/TLT.2015.2476806

Uto, M., & Ueno, M. (2018). Item response theory without restriction of equal interval scale for rater's score. In C. Penstein Rose, R. Martinez-Maldonado, H. Ulrich Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), Lecture notes in computer science, Vol. 10948. *Artificial Intelligence in Education. AIED 2018*. Springer. doi:10.1007/978-3-319-93846-2_68

Uto, M., & Ueno, M. (2020). A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika*, *47*(2), 469–496. doi:10.1007/s41237-020-00115-7

West-Pavlov, R. (Ed.). (2018). *The global south and literature (Cambridge Critical Concepts)*. Cambridge University Press. doi:10.1017/9781108231930

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283–306. doi:10.3102/10769986026003283

Minoru Nakayama is a professor at Information and Communications Engineering, Tokyo Institute of Technology, Japan. He completed the Master of Education program in 1985 from Tokyo Gakugei University and received a Doctor of Engineering degree from the Tokyo Institute of Technology in 1990. His research concerns human visual perception and educational technology.

Filippo Sciarrone is an associate professor at Universitas Mercatorum in Rome. He has been a Fellow Researcher, since 1994, at Roma Tre University, where he has been collaborating in research activities with the Artificial Intelligence Group and received a Ph.D. degree with a dissertation on user modeling. He has conducted several projects of industrial concern in the fields of user modeling and student modeling. He has led several research laboratories of private companies to produce algorithms and innovative systems for human resource management and for teaching-oriented recommendation systems.

Marco Temperini obtained a Ph.D. degree in computer science from Sapienza University of Rome, Rome, Italy, in 1992. He is currently an associate professor with the Department of Computer, Control, and Management Engineering at the same university. His research activity in the education arena is on adaptive e-learning, social and collaborative learning, game-based learning, automated assessment, and learning analytics. He has organized several international workshops and chaired a number of international conferences in the field of technology-enhanced learning. He has been a work-package leader and research unit coordinator for several European Union research projects.

Masaki Uto received a Ph.D. from the University of Electro-Communications, Tokyo, Japan, in 2013. He has been an associate professor at the University of Electro-Communications since 2020. He received the Best Paper Runner-up Award at AIED 2020. His research interests include educational and psychological measurement, Bayesian statistics, machine learning, and natural language processing.