Estimating Student's Performance Based on Item Response Theory in a MOOC Environment with Peer Assessment

Minoru Nakayama[1], Filippo Sciarrone[2], Masaki Uto[3], and Marco Temperini[4]

[1] Tokyo Institute of Technology, Meguro, Tokyo 152-8552, Japan,
nakayama@ict.e.titech.ac.jp
[2] Rome Tre University, Rome, Italy
sciarro@ing.uniroma3.it
[3] The University of Electro-Communications, Chofu, Tokyo , Japan
uto@ai.is.uec.ac.jp
[4] Sapienza University, Rome, Italy
marte@diag.uniroma1.it

Abstract. Peer Assessment is a powerful strategy to support educational activities and the consequent learners' success. Learning performance of participating is often estimated in a peer assessment setting using Item Response Theory. In this paper, a feasibility of estimating individual performance is examined for a simulated data set representing a MOOC environment, where one thousand students are supposed to perform a Peer Assessment session, where each peer assesses three other peers' work. For each student the modeling traits "ability", "consistency", and "strictness" are evaluated using Generalized Partial Credit Model, and the validity of such calculation is confirmed. While taking into consideration the limits of the synthetic sample production, this experiment provides an evidence of the possibility to predict learning performance in the large scale learning conditions of a MOOC.

Keywords: Peer Assessment, MOOC, Item Response Theory

1 Introduction

Massive Open Online Courses (MOOCs) have been developing for several years, with the aim of delivering learning contents to a numerous and worldwide audience [6]. They have well known and long studied problems with the dropout rates, which are in part due to the nature itself of these courses (still quite usually easy to enrol in and inexpensive), and in part for issues related to maintaining motivation and continuing engagement.

Peer Assessment (PA) is a powerful strategy to support educational activities and the consequent learners' success. With PA the learners can be exposed to different cognitive experiences: on the one hand they answer to a request (such as a question, or a task to be performed); on the other hand they are requested to assess other learners' work, being then involved in cognitive activities of higher level [3].

To encourage participant's learning activity, PA has been sometimes introduced in a MOOC [17] [1], although the reliability of the assessments, and in general the applicability of the strategy are still under verification.

PA has been proven reliable in applications where the teacher mediation and participation in the grading process are supported. In this approach the teacher grades a fraction of the learners' answers, and all the remaining answers have automated grading, based on the learners' models calibrated by the PA activity and by the available teacher's grades.

In particular, the learners modeling, deemed to represent the ability of the learner to answer a question, and the different ability to assess others' answers, has been studied in several research work. In [4] a Bayesian Network based modeling is defined, for the two learner's traits mentioned above, and a teacher mediated approach to PA is described. For this approach, involving a certain amount of teacher's grading (although just a fraction of the whole class of learners), an alternative has been proposed in [15], based on a modified version of K-NN algorithm [12], that is expected to be lighter from the computational point of view. This kind of teacher activity, though, is difficult to be granted if the number of students goes into the thousands, as is the case of MOOCs.

In [19] Item Response Theory (IRT) has been studied for similar purposes, conducting to a modeling of the learner's ability to assess. This method allows for a lighter computational cost of the process of modeling and automated assessment, but has the limit of having been experimented in too limited contexts (as far as the number of students is concerned).

The problem of having these approaches experimented on limited numbers of students is general, and in this paper we try and show a way of overcoming it, by the production of a "simulated MOOC" (i.e. a set of, say, 1000 simulated students, representing learners respecting a statistical distribution of their models' features).

This paper confirms the feasibility of determining the individual ability by using the IRT modelling technique applied to a huge PA data-set, based on the above simulation.

The remainder of the paper is structured as follows. In Section 2 the process of validation is shown, while in Section 3 the PA data-set building process is depicted. IRT is showed in Section 4. Our experimental results are in Section 5 while in Section 6 conclusions and future works are illustrated.

2 Process of validation for the IRT approach to PA

Basically, in the application of PA we are dealing with in this paper, we imagine that the PA system (PAS) will manage the data coming from a session of PA, and the grades provided by the teacher, to produce automated grading of the non-teacher-graded answers.

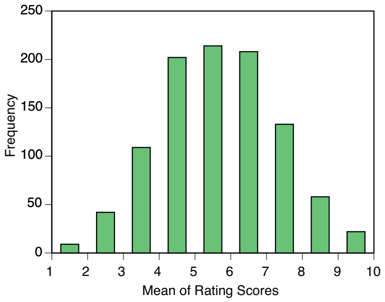

Fig. 1. Histogram of mean rating scores.

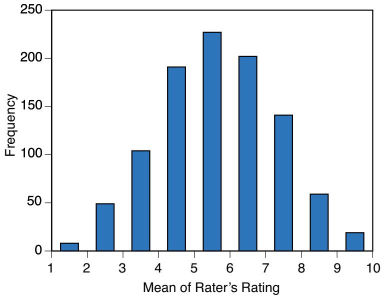

Fig. 2. Histogram of mean peer's ratings.

In a PAS there are three main data-sets of interest, described in the following paragraphs: the result of an experiment is in evaluating how well they are correlated, that is how well the PAS is able to compute valid learner's models and correct grading.

First there is the set of all the learners' models, describing the actual skills of the MOOC's members. We call this set Real (learners') Models (*RMs*). These data are generated according to a statistical distribution underlining them, and will be used to estimate the good behavior of the PAS.

Then there is the set of data coming from the PA activity (PA data-set). From the point of view of the data originated by a PA session, we can imagine that mainly two data-set are produced by Gaussian simulation, as follows: 1) the grades given by the learners to their peers (e.g. 3 grades for each peer, given and received); 2) the teacher's grade for each student's work. Notice that in real applications the data-set in 2) would contain only a fraction of the whole set of grades, whereas in our (simulated) experiments it covers the whole class (this is obtained by simulating a Gaussian distribution of the teacher's grades, so not involving real work from a real teacher).

Based on the PA data-Set, the PAS produces a new set of learners' grades. (Computed Models *CMs*). If the PAS is doing a good job, the new grades in *CMs* will be close to the ones in *RMs*. A more formal definition of the experiment we are going to present is as follows.

Definition 1. *An experiment of PAS consists of the following steps:*

1. *The RM data-set creation, based on a given statistical distribution;*
2. *The creation of the PA data-set, based on a given statistical distribution;*
3. *The application of the PAS algorithms by managing the creation and update of the grades in CM;*
4. *The computation of the grades provided by the PAS on those answers for which the teacher's grade in the PA data-set was not used;*

5. *Evaluation of the results, whereas the results can be the analysis of the correlation between the real and computed grades, that is between the teacher's real grades Vs. those computed by the PAS.*

In the following Section we discuss the creation steps of the experiment. In Section 4 we show the application of IRT with the aim of validating the use of IRT to a (simulated) MOOC context.

In Section 4 we will use a modeling consistent with the features of IRT, and in particular with the features of a *Generalized Partial Credit Model* with Rater Parameters. By this modeling technique, the learner j is represented by the following features:

- θ_j, latent ability of j
- α_j, consistency of rater j
- β_{jk}, strictness of the rater j for rating category k

3 The Peer Assessment Data-Set

In this section we briefly report on the process of creation of the *RM* and *PA data-set*, provided by our simulation system [16].

3.1 The Data Generation Procedure

First data to be generated is the PA data-set, based on a specific graphical interface used by the experimenter (typically a teacher). The interface allows to generate the grades given by each peer to her/his n peers (in our use of the simulation system $n = 3$).

These grades are are not actually given directly by the teacher, rather they are generated randomly, according to the classic bell-shaped Gauss distribution, which has proven to be very reliable in representing grades distribution in a learning context. For such distribution, the parameters μ (the value of the mean), and σ^2 (variance) are set by the experimenter. In our case $\mu = 5.5$ and $\sigma^2 = 1.63$

The student models are then implicitly defined based on the PA data-set, as we suppose that each student in the experiment is behaving accordingly to her/his model.

3.2 The Features of the Generated Data

The simulated data-set consists of 1000 participants, where three of them assess another participant's performance using a 10 points scale, i.e. a participant's performance i, r_i, is computed by the grades given by other three participants such as $i + 1, i + 2, i + 3$, here $i = 1, \ldots, 1000$.

Simple statistics of generated data are regulated by the algorithm. Both means of rating scores and peer's ratings are summarized in Figures 1 and 2. Also, the relationship between two means is illustrated in Figure 3.

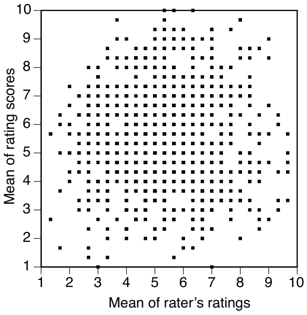

Fig. 3. Scatter gram between peer's ratings and rating scores for each participant

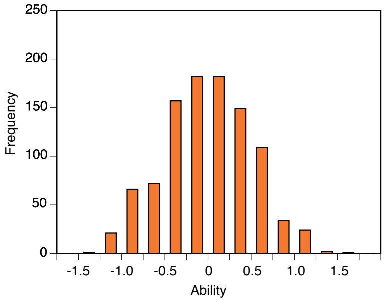

Fig. 4. Histogram of calculated ability of participants.

4 The Item Response Theory

This study attempt to apply IRT [10], a test theory based on mathematical models, to the PA data. IRT models provide an item response function that specifies the probability of a response to a given test item as a function of latent participant's ability and the item's characteristics such as difficulty and discrimination. IRT enables to estimate participant ability while considering item characteristics.

IRT also has the following advantages: i) participant responses to different test items can be assessed on the same scale; ii) missing data can be easily estimated.

IRT has traditionally been applied to test items for which responses can be scored as correct or incorrect. In recent years, however, there have been attempts to apply polytomous IRT models to performance assessments, including PA.

4.1 The Generalized Partial Credit Model

A representative polytomous IRT model is the generalized Partial Credit Model (GPCM) [13]. The GPCM gives the probability that participant j receives score k for test item i as

$$P_{ijk} = \frac{\exp \sum_{m=1}^{k} [\alpha_i(\theta_j - \beta_{im})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_i(\theta_j - \beta_{im})]} \ , \tag{1}$$

where:
α_i is a discrimination parameter for item i, β_{ik} is a step difficulty parameter denoting difficulty of transition between scores $k-1$ and k in the item, and θ_j is the latent ability of participant j. Here, $\beta_{i1} = 0$ for each i is given for model identification.

4.2 The GPCM with Rater Parameters

As described in the Introduction, this study applies IRT to PA data comprising participants × peer-raters. However, the traditional IRT models introduced above are not directly applicable to such data. To address this problem, many IRT models that incorporate rater characteristic parameters have been proposed [20].

This study introduces a state-of-the-art GPCM model incorporating rater parameters [21]. This model provides the probability that peer-rater r assigns score k to participant j's performance for item (or performance task) i as

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - \beta_{rk})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - \beta_{rk})]} \ , \tag{2}$$

where, α_r reflects the consistency of rater r, β_{rk} is a step difficulty parameter denoting difficulty of transition between scores $k-1$ and k in the rater r. β_r is represents the strictness of rater r, and β_i is the difficulty of item i. Here, $\prod_i \alpha_i = 1$, $\sum_i \beta_i = 0$, $d_{r1} = 0$, and $\sum_{k=2}^{K} d_{rk} = 0$ are given for model identification.

This study applies this model to the PA data described in Section 3. It is noteworthy that, in the data, the number of performance tasks is fixed to one. In this case, α_i and β_i are ignorable because the model identification constraints restricts the value of $\alpha_{i=1} = 1$ and $\beta_{i=1} = 0$.

4.3 The Parameter Estimation

To estimate IRT model parameters, marginal maximum likelihood estimation using an EM algorithm has been commonly used [2]. However, for complex models like that used in this study, EAP estimation, a form of Bayesian estimation, is known to provide more robust estimations [5]. EAP estimates are calculated as the expected value of the marginal posterior distribution of each parameter. For the EAP estimation, Markov Chain Monte Carlo (MCMC), a random sampling–based estimation method, is generally used [5, 18].

The Metropolis-Hastings-within-Gibbs sampling algorithm has been used as a MCMC algorithm for IRT models [14]. The algorithm is simple and easy to implement, but it requires long times to converge to the target distribution [8]. The Hamiltonian Monte Carlo (HMC) is an alternative MCMC algorithm with high efficiency. In recent years, the No-U-Turn (NUT) sampler [8], an extension of HMC that eliminates hand-tuned parameters, has been proposed. Because the "Stan" software package makes implementation of a NUT-based HMC easy, this algorithm has recently been used for parameter estimations in various statistical models, including IRT models [11, 9].

For the aforesaid reasons, we use a NUT-based MCMC algorithm for parameter estimations. We calculate EAP estimates as the mean of parameter samples obtained from 500 to 1,000 periods of three independent MCMC chains. Furthermore, the standard normal distribution is used as the prior distributions.

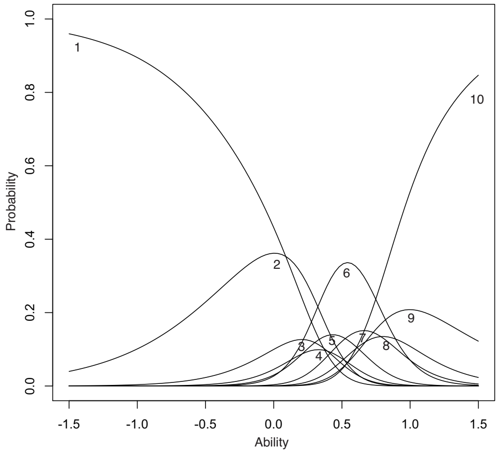

Fig. 5. Curves of item category response function (ICRF) for the 4th rater.

5 Results

We estimated the IRT parameters from the PA data described in Section 3.

Table 1. Simple statistics of ratings and estimated parameters.

variable	N	mean	STD
Rating scores	1000	5.50	1.63
Peer's rating	1000	5.50	1.66
Ability	1000	0.00	0.51
Consistency	1000	0.90	0.46
Strictness	1000	0.00	0.48

Though each rater gave scores to three participants only, such as rates "1", "4" and "6", the MCMC algorithm estimated the IRT parameters. The MCMC was run using 2.5 GHz 14 core Intel Xeon W processor. The calculation time was 1628.57 seconds. We confirmed the Gelman–Rubin statistic $\hat{R}$ [7], which is generally used as a convergence diagnostic. Values for these statistics were less than 1.1 for all parameters, indicating that the MCMC runs converged.

Table 1 shows the overall simple statistics for learner's model features ability, rater consistency and strictness. Furthermore, as an example, Figure 5 depicts the curves of item category response function (IRCs) for the 4-th rater. In the Figure, the horizontal axis represents the latent ability θ, while the vertical axis shows the response probability for each rating category. The figure indicates a characteristic of overall rating behaviors.

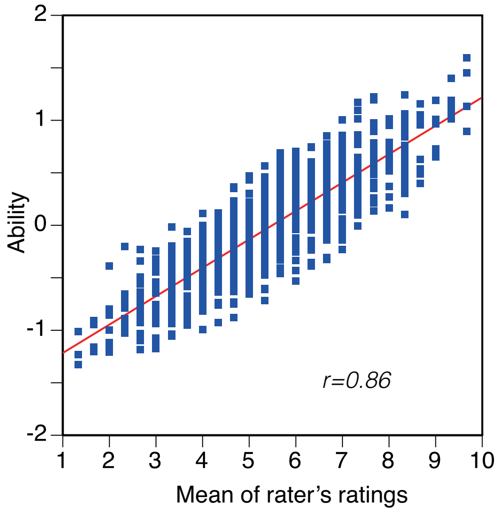

Fig. 6. Relationship between mean peer's ratings and the estimated ability.

To illustrate a validation of estimated ability using our model, the relationship between the ability and means of rater's ratings is summarized in Figure 6. The Figure shows a strong correlation, with $r = 0.86$. In the Figure, a regression line is overlapped to the scatter-plot. Since rating behaviour affects to estimate the ability, some deviations for the ability are observed.

In order to examine the rating behaviours, the relationships between rating score and consistency, and rating score and strictness are summarized in Figure 7 and 8 respectively. The consistency is again deviating over the rating scores and the behaviour may be controlled by the data generation procedure. The strictness decreases with rating score, the correlation coefficient is $r = -0.89$. The strictness indicates severeness of rating so that there is negative relationship between them. In addition to the relationship, the data generation may influence some deviation of the strictness which are observed in the Figure.

6 Conclusions and Future Work

This paper examined a possibility of estimating student's performance which was based on a simulated data set using the IRT. The simulated data-set is generated with 1000 students and with PA based on the requirement of having 3-peer's grading per each student, where a $1 - 10$ point scale is used for the grades. The performance was estimated by using the GPCM and some detailed parameters were calculated with EAP estimation and MCMC techniques.

As a result, we had the opportunity to estimate values of ability, consistency and strictness for each participant, in a learning context characterized by one thousand students.

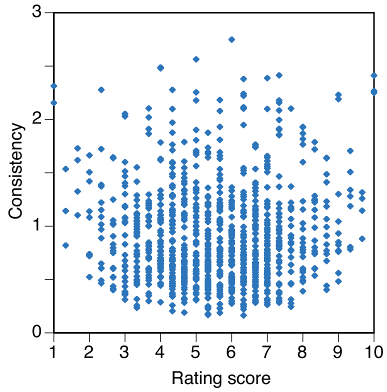

Fig. 7. Relationship between mean rating scores and the estimated consistency.

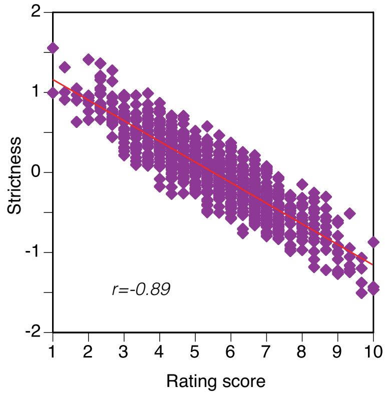

Fig. 8. Relationship between mean rating scores and the estimated strictness.

We modeled the students by means of the above mentioned parameters, and the feasibility of the estimation for a large data set was examined, confirming the validity of the simulated approach for the trial.

In particular, the limited experimentation we presented allowed us to plan a more extensive use of simulation of PAS in a MOOC context, and the enrichment of the PAS capabilities by means of IRT. Such future work will have the intent of checking on PAS aspects, that can have influence on the PAS performance, and see how their improvements could make the PAS better. By PAS aspects above, we mean, for instance, the number of peers in the class, the number of peer assessments required by each peer, the applied rating scale, the tuning of teacher's grading, by recommendation of what peer's work would be more beneficial to grade by the teacher, in order for the PAS to produce a better automated grading, and in general the algorithms to actually computing the automated grading in a PA session.

Acknowledgement

This research was partially supported by the Japan Society for the Promotion of Science (JSPS), KAKEN (17H00825).

References

1. Alcarria, R., Bordel, B., deAndrés, D.M., Robles, T.: Enhanced peer assessment in MOOC evaluation through assignment and review analysis. International Journal of Emerging Technologies in Learning 13(1), 206–219 (2018)
2. Baker, F., Kim, S.H.: Item Response Theory: Parameter Estimation Techniques. Statistics, textbooks and monographs, Marcel Dekker (2004)

3. Bloom, B.S.: Taxonomy of Educational Objectives. David McKay Company Inc., New York, USA (1964)
4. De Marsico, M., Sciarrone, F., Sterbini, A., Temperini, M.: Supporting mediated peer-evaluation to grade answers to open-ended questions. EURASIA Journal of Mathematics, Science and Technology Education 13(4), 1085–1106 (2017)
5. Fox, J.P.: Bayesian item response modeling: Theory and applications. Springer (2010)
6. de Freitas, S.I., Morgan, J., Gibson, D.: 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 (2015)
7. Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Statist. Sci. 7(4), 457–472 (1992)
8. Hoffman, M.D., Gelman, A.: The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 15, 1593–1623 (2014)
9. Jiang, Z., Carter, R.: Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via Stan. Behavior Research Methods 51(2), 651–662 (2019)
10. Lord, F.: Applications of item response theory to practical testing problems. Erlbaum Associates (1980)
11. Luo, Y., Jiao, H.: Using the Stan program for Bayesian item response theory. Educational and Psychological Measurement 78(3), 384–408 (2018)
12. Mitchell, T.M.: Machine Learning, 1st edn. David McKay, New York, USA (1997)
13. Muraki, E.: A generalized partial credit model. In: van der Linden, W.J., Hambleton, R.K. (eds.) Handbook of Modern Item Response Theory, pp. 153–164. Springer (1997)
14. Patz, R.J., Junker, B.: Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics 24, 342–366 (1999)
15. Sciarrone, F., Temperini, M.: K-openanswer: a simulation environment to analyze the dynamics of massive open online courses in smart cities. Soft Computing. In Press. (2020)
16. Sciarrone, F., Temperini, M.: Simulating massive open on-line courses dynamics. In: Proceedings of iTHET 2019. pp. 1–9. Magdeburg, Germany (2019)
17. Sun, D.L., Harris, N., Walther, G., Baiocchi, M.: Peer assessment enhances student learning: The results of a matched randmized crossover experiment in a college statistics class. PLoS ONE 10(12), 1–7 (2015)
18. Uto, M.: Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In: Proc. International Conference on Artificial Intelligence in Education. pp. 494–506 (2019)
19. Uto, M., Ueno, M.: Item response theory for peer assessment. IEEE Transaction of Learning Technology 9(2), 157–170 (2016)
20. Uto, M., Ueno, M.: Empirical comparison of item response theory models with rater's parameters. Heliyon 4, 1–32 (2018)
21. Uto, M., Ueno, M.: Item response theory without restriction of equal interval scale for rater's score. In: Proc. International Conference on Artificial Intelligence in Education. pp. 363–368 (2018)