So, H. J. et al. (Eds.) (2020). Proceedings of the 28th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

A warm-up for adaptive online learning environments – the Elo rating approach for assessing the cold start problem

Maciej PANKIEWICZ
Institute of Information Technology, Warsaw University of Life Sciences, Poland
maciej_pankiewicz@sggw.pl

Abstract: The aim of this study is to present and evaluate the Elo rating algorithm as a tool for assessing the task difficulty in terms of the so-called "cold-start" problem – during the initial phase of the introduction of the adaptive system to the public. This analysis has been performed on the real data originating from the online programming course available on the RunCode platform: the online learning environment with multiple attempts allowed and feedback provided after every attempt. There have been 50055 submissions on 76 tasks uploaded by 299 RunCode users. It has been found that the Elo rating algorithm achieves the correlation of 0.702 with the reference values already for the sample size of n = 5, and the correlation of 0.905 for the sample size of n = 50. The Elo algorithm outperforms the Proportion Correct method for small sample sizes and may be a more reasonable choice as a simple method for task difficulty estimation during the initial phase of introducing the adaptive system to the public.

Keywords: Difficulty estimation, Elo rating, proportion correct, automated assessment, programming course

1. Introduction

The problem of the task difficulty estimation – if the amount of data is sufficient – may be assessed with several models e.g. originating from the Computerized Adaptive Testing (CAT) domain. The Item Response Theory (IRT) provides a range of well-established methods for estimating task difficulty (Rasch, 1960, 1966, 1977) that have been not only utilized in educational (Scheerens et al., 2006), but also medical (Christensen et al., 2013) or marketing applications (Bechtel, 1985). There is however a fundamental limitation of these methods if considering the initial phase of the introduction of the adaptive online learning environment to the public – the requirement of large calibration samples.

There have been several alternative methods for task difficulty estimation examined by the research community in an educational context, e.g. the Elo rating algorithm (Antal, 2013; Klinkenberg et al., 2011; Pankiewicz, 2020a; Pankiewicz & Bator, 2019; Pelánek et al., 2017; Wauters et al., 2012), proportion correct (Antal, 2013; Wauters et al., 2012), or learner feedback method (Chen et al., 2005; Wauters et al., 2012). It has been observed that – for sufficiently big datasets – all of the above-mentioned methods may deliver estimations characterized by reasonably high accuracy. The performance of these methods if the dataset is small has not been extensively researched, especially in the context of environments with the formative assessment approach – where multiple attempts are allowed and feedback is provided after every attempt.

The so called "cold start" problem (Schein et al., 2002) in adaptive learning environments refers to the situation, where little is known about ability of system users and/or difficulty level of items available in the system. Several approaches originating from the machine learning domain (Pliakos et al., 2019; Wei et al., 2017) have been proposed to address this issue in educational context. Recently, also the Elo-based method has been introduced (Park et al., 2019), that integrates the explanatory IRT model. However, if the system is fresh, estimation algorithms are untrained and the introduction of the alternative estimation method is temporary, involvement of sophisticated methods may not be in the focus for the implementing team. Therefore, the question remains if (and to what extent) these less complex methods characterized by low implementation and computational requirements may support the initial phase of the system deployment in adaptive learning environments with the formative assessment approach.

The focus of this research is therefore to examine, to what extent the original Elo rating algorithm may be an appropriate choice for the task difficulty estimation in terms of the so-called “cold start” problem – during the initial phase of the introduction of the adaptive system to the public. The source of the data has been the item-based programming course available on the RunCode online learning environment (Pankiewicz, 2020b). Additionally the Elo estimations have been compared with another simple measure: the Proportion Correct – a method that has been previously found to deliver more accurate estimations than Elo rating algorithm (Antal, 2013; Wauters et al., 2012).

2. Estimating difficulty of an assignment

2.1 Elo rating algorithm

The Elo rating system (Elo, 1978) has been developed for the purpose of measuring strength of players in chess tournaments. The aim of the algorithm is to calculate players’ rating change after every game. That change depends on outcomes of tournament games. Every player is assigned a rating that is usually a number between 1000 and 3000 that is a subject to change after every game. New rating is calculated by a formula:

$$R_n = R + K(O - P) \tag{1}$$

Where: R_n is the new value of the rating, R – the actual rating, O – game outcome (1 – win, 0 – loss), P – probability of winning the game and constant K – the value for chess tournaments is often 32. The probability of winning P is given as:

$$P = \frac{1}{1 + 10^{\frac{R_o - R_p}{400}}} \tag{2}$$

Where R_p is the rating of a player and R_o is the rating of the opponent. In the context of an online learning environment, we consider a tournament game to be a single submission of a solution, a player – a learner that submits the solution and opponent – a task.

There are three possible outcomes of the chess game (win, loose, draw), but in the context of a learning environment we only consider two outcomes: learner wins if the submission receives the maximum score or learner loses if the submission does not receive maximum score.

2.2 Proportion correct

Proportion correct (PC) – the percentage of correct answers is another simple measure to assess item difficulty calculated for every *i*-th item as:

$$\widehat{b_i} = 1 - \frac{n_i}{N_i} \tag{3}$$

Where n_i is the number of correct attempts and N_i the number of total attempts on the *i*-th task.

PC is calculated as the number of learners that solved the task divided by the total number of learners therefore the more learners solved the task in total, the lower the difficulty of that task. According to (Wauters et al., 2012), proportion correct may generate accurate estimations if administered to already 200-250 learners. The accuracy of the method is very high according to several research (Antal, 2013; Wauters et al., 2012).

3. Methodology

3.1 Data

The data originates from the RunCode online learning environment available at https://runcodeapp.com. The online course that supports automated verification of programming tasks has been made available on the RunCode platform as an additional tool in the *Introduction to programming* course – a mandatory course for the first-semester computer science students at the Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences. There have been

50055 attempts on 76 tasks recorded on the RunCode platform during two editions of the course: 2017/2018 and 2018/2019. There were 299 students in total that used the system during those two editions. Usage of the platform was not mandatory, however majority of students actively participated.

Two methods for estimating task difficulty have been compared in this study: the Elo rating algorithm and Proportion Correct. In order to compare the accuracy of these methods in terms of the cold start problem, reference values of task difficulty have been obtained on the full dataset by the means of the IRT graded response model (Samejima, 1969). The graded response model is suitable for modelling polytomous response data and has been already introduced e.g. for the purpose of knowledge assessment on open-ended tasks with multiple attempts allowed (Attali, 2011).

3.2 Sample size computations

The comparison of difficulty estimation methods is presented in regards to the number of attempts. For the *i*-th attempt, calculations include the cumulative number of all trials on tasks that do not exceeded *i* attempts recorded for a single learner on the task. The sample limited by *i* attempts for a particular learner is selected from all attempts recorded in the data set. No further attempt is contained in the analysis after the attempt that received maximum score. If a learner received a maximum score and submitted another solution, it was not considered in the analysis.

As the outcome of analyses we considered a list of tasks ordered by a difficulty rating. In order to compare difficulty estimations provided by the discussed methods the Pearson’s correlation coefficient has been used.

In order to assess the cold start problem, random samples of sizes n = 5, 10, 20 and 50 learners have been drawn from the data set with assurance that at least one attempt has been recorded for every task. The selection procedure has been repeated 1000 times for each sample size. The analysis contains all tasks on which there was at least one attempt, limited to the number of *i* attempts for every learner. Correlation has been calculated for each sample drawn. Then, the median of obtained correlations has been calculated to acquire the “typical” value. Median has been chosen in order to limit the impact of outliers on final results.

4. Results

This analysis is aimed at the evaluation of the Elo rating algorithm in terms of the so-called “cold-start” problem within adaptive online learning environments. Results of the study show that Elo rating algorithm may be a good choice for computing task difficulty estimations during the initial stage of the introduction of the adaptive system to the public, if there is no sufficient amount of data in order to utilize more accurate methods. It has been compared to the Proportion Correct (PC) method – another simple difficulty estimation measure. The Elo method achieves distinctly higher correlation values with the reference data than the PC method for small sample sizes. With increasing size of the sample, the difference between the Elo and Proportion Correct decreases.

The data originates from the programming course available at the RunCode online learning platform used by users of varying programming skills. The course has been made available to the students of the *Introduction to programming* – a mandatory course at the *Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences*. Before joining the course students answered in a survey on their self-evaluation of programming skills. More than a half of students declared to have low or very low level of programming knowledge before joining the course.

The average number of unsuccessful attempts on tasks preceding the successful trial was high. Despite the fact that the average difficulty level may be perceived as high, the engagement of platform users was surprisingly high: students did not quickly resign and mostly uploaded another solution.

For the purpose of clarity, results have been presented for the first seven attempts (ca. 75% of all samples). This limitation is reasonable, as the effects visible on the first seven attempts are in general also reflected in the remaining data (e.g. dropout) with the long tail of even more than 50 attempts on a task. The detailed analysis of the submission data has been presented on Table 1.

Table 1. *The number of correct and incorrect attempts on assignments. The Total column is the cumulative sum of attempts. The Dropout column is the percentage of students that resigned to take another attempt.*

Attempts	*Incorrect*	*Correct*	*Total*	*Dropout*
1	8259	5269	13528	-
2	5623	2389	21540	0.030
3	4045	1244	26829	0.059
4	2950	842	30621	0.063
5	2259	493	33373	0.067
6	1766	342	35481	0.067
7	1396	237	37114	0.075

The sampling procedure presented in the section 3.2 has been performed on the following sizes of the sample: n = 5, 10, 20 and 50. Results of the analysis – Pearson's correlation values between reference values and the Elo and PC difficulty estimations for the first seven attempts – have been presented on the Figure 1.

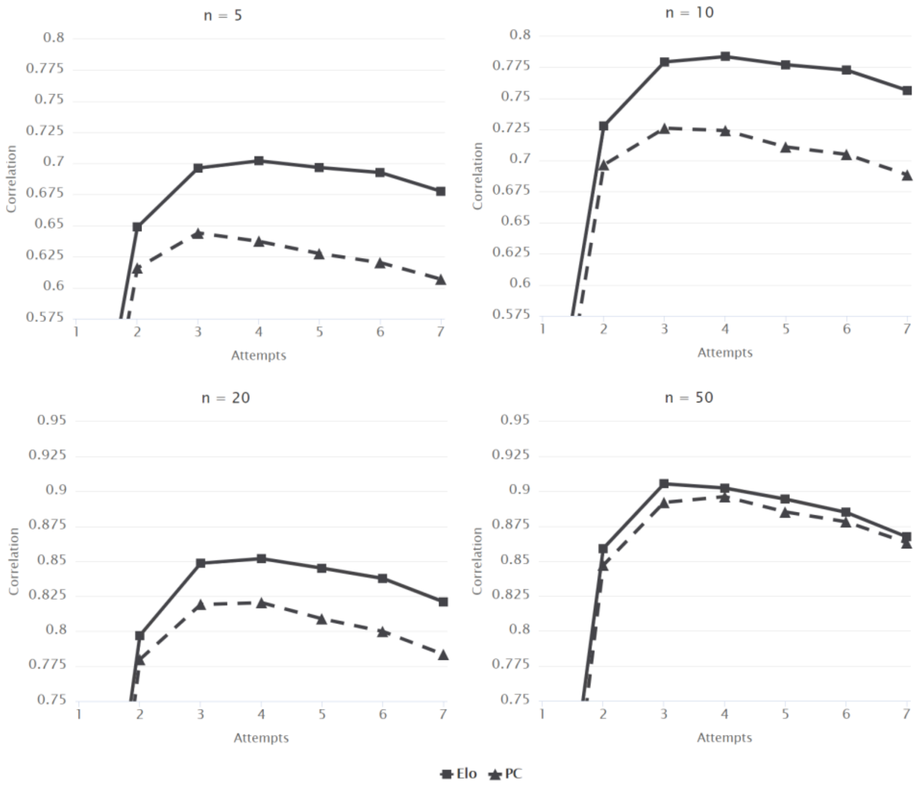

Figure 1. Correlation of the item difficulty estimates with the reference values for the range of attempts between 1 and 7 and sample size of 5, 10, 20 and 50. Estimation methods: Elo rating algorithm, proportion correct (PC).

The highest value of the correlation for the size of the sample n = 5 between the Elo estimation and the reference values: $\text{cor}_{\text{ELO5S}} = 0.702$, PC: $\text{cor}_{\text{PC5S}} = 0.644$. For the size of the sample n = 10, the highest observed correlation value $\text{cor}_{\text{ELO10S}} = 0.784$, PC: $\text{cor}_{\text{PC10S}} = 0.726$. For n = 20, the highest value $\text{cor}_{\text{ELO20S}} = 0.852$, PC: $\text{cor}_{\text{PC20S}} = 0.821$. For n = 50, the highest value $\text{cor}_{\text{ELO50S}} = 0.905$, PC: $\text{cor}_{\text{PC50S}} = 0.896$. For all of analyzed sizes of the sample, the Elo rating algorithm achieved higher values of correlation than the proportion correct method (Table 2).

Table 2. *Correlation of the item difficulty estimates with the reference values for the range of attempts between 1 and 7 and sample size of 5, 10, 20 and 50. Estimation methods: Elo rating algorithm, proportion correct (PC).*

	Size = 5		*Size = 10*		*Size = 20*		*Size = 50*	
attempts	*Elo*	*PC*	*Elo*	*PC*	*Elo*	*PC*	*Elo*	*PC*
1	0,378	0.343	0.433	0.395	0.484	0.439	0.533	0.487
2	0.649	0.616	0.728	0.697	0.797	0.780	0.859	0.847
3	0.696	0.644	0.779	0.726	0.849	0.819	0.905	0.892
4	0.702	0.637	0.784	0.724	0.852	0.821	0.902	0.896
5	0.697	0.627	0.777	0.711	0.845	0.809	0.894	0.885
6	0.693	0.620	0.773	0.705	0.838	0.800	0.885	0.878
7	0.678	0.607	0.756	0.688	0.821	0.783	0.867	0.863

5. Summary and discussion

The aim of this study was to present the Elo rating algorithm as a tool for assessing the task difficulty in terms of the so-called “cold-start” problem. This analysis has been performed on the real data originating from the online item-based *Introduction to programming* course available on the RunCode platform: the online learning environment where multiple attempts were allowed and feedback was provided after every attempt. The so-called “cold start” problem refers to the initial stage of the introduction of an adaptive learning system to the public, where little is known about system users and/or available items. Until a sufficient amount of data is gathered and item bank calibration may be performed, usage of methods e.g. originating from the area of the Computerized Adaptive Testing domain becomes a hurdle. Therefore alternative, simpler methods of difficulty estimations, such as Elo rating algorithm may be considered as a temporary (or sometimes permanent) solution. Results of this study showed that the Elo rating algorithm achieved the correlation of 0.702 with the reference difficulty estimation values obtained by the means of the IRT graded response model for the size of the sample n = 5. For the size of the sample n = 10: cor. = 0.784, for n = 20: cor. = 0.852, and for n = 50: cor. = 0.905. Estimations obtained by the Elo method outperform values calculated by the proportion correct measure for sizes of the sample n = 5, 10 and 20. As the size of the sample increases, difference between estimations calculated by the Elo rating algorithm and Proportion Correct method decreases.

The first conclusion drawn from results of this study is that the introduction of the Elo rating algorithm for the purpose of assessing task difficulty at the initial stage of the introduction of the adaptive system to the public may be a reasonable choice. Already for the size of the sample as small as n = 5, the method achieves reasonably high correlation of 0.702. For larger size of the sample n = 50, the Elo rating algorithm achieves the correlation value of 0.905, however Proportion Correct achieves comparable value: 0.896. Both methods are characterized by low computational requirements, and – compared to e.g. expert rating or learner feedback – do not require additional human engagement. There is also an additional benefit: the complexity of the method is low and therefore its implementation may be easily carried out.

The second conclusion is that the Elo rating algorithm performs better than the Proportion Correct method for smaller sizes of the sample. The Elo method quickly “learns” from user submission results and therefore much quicker adjusts its difficulty rating. As the number of recorded submissions increases, this benefit of the Elo method is less visible in the outcome and estimations of both PC and Elo methods become similar. There is no direct comparison to be established with previous research, as it has been mainly focused on the summative assessment: in the study of (Wauters et al., 2012) both the Proportion Correct and Elo method similarly estimate reference values for small size of the sample with small advantage of the PC method, and in the study of (Antal, 2013) the Proportion Correct method delivers more accurate estimations.

Main limitations of this study refer to the aspect of high engagement observed within analyzed group of platform users. The observed dropout rate has been very low and may be a result of implemented gamification elements, but also may result from the fact that system users were university’s computer science students. Although the usage of the platform was not mandatory, and results obtained within the platform did not impact final grade, the motivation of students may have been higher than of an average group of people interested in gaining experience in programming. It is more probable that these observations will be replicable in university settings, than on any publicly available online learning platform.

References

Antal, M. (2013). On the use of elo rating for adaptive assessment. *Studia Universitatis Babes-Bolyai, Informatica*, *58*(1). https://www.researchgate.net/publication/301635151

Attali, Y. (2011). Immediate feedback and opportunity to revise answers: Application of a graded response IRT model. *Applied Psychological Measurement*, *35*(6), 472–479. https://doi.org/10.1177/0146621610381755

Bechtel, G. G. (1985). Generalizing the Rasch Model for Consumer Rating Scales. *Marketing Science*, *4*(1), 62–73. https://doi.org/10.1287/mksc.4.1.62

Chen, C. M., Lee, H. M., & Chen, Y. H. (2005). Personalized e-learning system using Item Response Theory. *Computers and Education*, *44*(3), 237–255. https://doi.org/10.1016/j.compedu.2004.01.006

Christensen, K. B., Kreiner, S., & Mesbah, M. (2013). Rasch Models in Health. In *Rasch Models in Health*. John Wiley and Sons. https://doi.org/10.1002/9781118574454

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Klinkenberg, S., Straatemeier, M., & Van Der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education*, *57*(2), 1813–1824. https://doi.org/10.1016/j.compedu.2011.02.003

Pankiewicz, M. (2020a). Measuring task difficulty for online learning environments where multiple attempts are allowed — the Elo rating algorithm approach. *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020*, 648–652.

Pankiewicz, M. (2020b). Move in the Right Direction: Impacting Students' Engagement With Gamification in a Programming Course. *Proceedings of EdMedia + Innovate Learning*, 1180–1185. https://www.learntechlib.org/primary/p/217434/

Pankiewicz, M., & Bator, M. (2019). Elo Rating Algorithm for the Purpose of Measuring Task Difficulty in Online Learning Environments. *E-Mentor*, *82*(5), 43–51. https://doi.org/10.15219/em82.1444

Park, J. Y., Joo, S. H., Cornillie, F., van der Maas, H. L. J., & Van den Noortgate, W. (2019). An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments. *Behavior Research Methods*, *51*(2), 895–909. https://doi.org/10.3758/s13428-018-1166-9

Pelánek, R., Papoušek, J., Řihák, J., Stanislav, V., & Nižnan, J. (2017). Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction*, *27*(1), 89–118. https://doi.org/10.1007/s11257-016-9185-7

Pliakos, K., Joo, S. H., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers and Education*, *137*, 91–103. https://doi.org/10.1016/j.compedu.2019.04.009

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests.*

Rasch, G. (1966). An Item Analysis Which Takes Individual Differences Into Account. *British Journal of Mathematical and Statistical Psychology*, *19*(1), 49–57. https://doi.org/10.1111/j.2044-8317.1966.tb00354.x

Rasch, G. (1977). On Specific Objectivity. An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements in Symposium on Scientific Objectivity, Vedbaek, Mau 14-16, 1976. *Danish Year-Book of Philosophy Kobenhavn*, *14*, 58–94.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. In *Psychometrika* (Vol. 34). https://doi.org/https://doi.org/10.1007/BF03372160

Scheerens, J., Glas, C., & Thomas, S. M. (2006). Educational evaluation, assessment and monitoring: A systematic approach. In *Educational Evaluation, Assessment and Monitoring: A Systematic Approach*. Taylor & Francis Group. https://doi.org/10.4324/9780203971055

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '02*, 253. https://doi.org/10.1145/564376.564421

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers and Education*, *58*(4), 1183–1193. https://doi.org/10.1016/j.compedu.2011.11.020

Wei, J., He, J., Chen, K., Zhou, Y., & Tang, Z. (2017). Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2016.09.040