

# A Self-adaptive Mechanism for Serious Quiz Games

Michael Striewe<sup>a</sup>

*University of Duisburg-Essen, Essen, Germany*

**Keywords:** Serious Quiz Games, Self-Adaptation, Quiz Difficulty.

**Abstract:** Serious quiz games can provide an entertaining way to assess knowledge and promote learning. Motivation of the players can be increased by presenting quiz questions in the order of increasing difficulty or by presenting many different quiz questions during subsequent quiz sessions. However, maintaining a question set to adjust difficulty ratings for the target audience and to avoid early repetitions of questions can be a challenging task that requires constant effort. The paper presents a self-adaptive mechanism that is able to solve both tasks without collecting any explicit information about the anonymous target audience. The evaluation demonstrates that a quiz with about 600 questions on 15 difficulty levels has been maintained successfully over several years and more than 100'000 quiz sessions with minimal manual intervention.

## 1 INTRODUCTION

Serious quiz games can be an entertaining and motivating tool for self-assessment. They can promote learning in various educational scenarios or help teachers and students to track progress. Although empirical results reveal some downsides or limited effects, a general positive effect can be confirmed in several studies (Heitmann et al., 2021; Becker-Blease and Bostwick, 2016; Simon-Campbell and Phelan, 2016; Wang, 2008; Wickline and Spektor, 2011). Serious quiz games are thus also used in informal learning settings, where the joy of playing the quiz game is equally important to the learning gain.


A recent survey on using KAHOOT!, a game-based learning platform, reveals that a particular challenge for teachers is to get the difficulty level of questions and answers right (Wang and Tahir, 2020). Negative effects e. g. on motivation can be expected both for presenting too easy questions to advanced players as well as for presenting too hard questions for beginners. Another negative effect on motivation could possibly be observed if questions are shown repeatedly to the same player within a short time frame. If little is known about the target audience and thus it cannot be foreseen which difficulty levels will be used most, constant effort is required to adjust the size of question pools for each level to the actual demands.

Both aspects are particularly important in informal settings: A quiz game that is not motivating will not

be used as there is no additional extrinsic motivation in these settings. In informal settings there may also be less resources available for maintaining and calibrating the question pool. Finally, there may also little be known about the level of knowledge within the target audience. Hence, sophisticated methods for generating quizzes (like e. g. proposed by (Lin, 2020)) are hardly applicable in these contexts.

This paper thus proposes a self-adaptive mechanism to solve these challenges: The mechanism automatically maintains a difficulty rating for each question and adjusts the number of questions used per level based on data from previous quiz sessions. Different to other approaches like IRT or Elo-Ratings (Mangaroska et al., 2019; Park et al., 2019), it does not maintain a rating for each player and is hence suitable for anonymous quiz games in informal settings. The quiz mechanism has been implemented on a public website and been used in more than 100'000 quiz sessions. Based on these data and a close inspection of all quiz sessions from the year 2021, the paper evaluates whether both the difficulty rating of the quiz questions and the behaviour of players in recent quiz sessions show the expected characteristics. The evaluation aims to find out whether the mechanism is indeed able to solve the aforementioned challenges.

The remainder of the paper is organized as follows: Section 2 provides a detailed description of the quiz mechanism and the reasoning for its design. Section 3 provides details on how an actual implementation of the quiz mechanism is used on a public web-

<sup>a</sup>  <https://orcid.org/0000-0001-8866-6971>

site. Section 4 provides usage data for the quiz implementation and discusses whether the behaviour is as expected. Section 5 concludes the paper.

## 2 QUIZ MECHANISM

The quiz mechanism is intended to work with quizzes that associate a difficulty level and a category or topic with each quiz question. The assumed main use case is to serve quiz sessions that present questions of increasing difficulty level as long as questions are answered correctly and the maximum level is not yet reached. To ensure large coverage of a domain, at most one question per category or topic is to be used during one session. Categories can also be used if there are questions that are mutually exclusive in the sense that one question text may directly reveal the answer to another question. Nevertheless, the use of categories has no further implications for the quiz mechanism and can thus be considered optional.

The quiz mechanism could also be used for other forms of quizzes, such as serving sessions of fixed length, where each correct answer is followed by a harder question and each wrong answer by an easier one. Finally, the quiz mechanism is agnostic to the form of quiz question as long as answers are strictly considered either right or wrong.

The quiz mechanism operates on a simple infrastructure that can be implemented in form of two data tables: One data table contains the quiz questions and particularly stores the category or topic for each question and a difficulty rating in the range from 0 to 1, where higher numbers denote easier questions. The second data table contains an entry for each difficulty level and stores the number of questions delivered so far for that level.

### 2.1 Adapting the Question Rating

New questions can be inserted into the question set with any initial difficulty rating in the range of 0 to 1. It is advisable to make an educated guess on an appropriate rating, but it is not crucial to be very precise. The difficulty rating will be updated each time the question is answered according to the following rule:

$$rating_{new} = \begin{cases} rating_{old} * 0.99 + 0.01 & \text{if correct,} \\ rating_{old} * 0.99 & \text{if wrong} \end{cases}$$

Hence, a question with rating 0.5 will get a new rating of 0.505 if answered correctly and a new rating of 0.495 if answered wrong. The formula implies that

a more recent answer has a larger impact on the rating than an older one. Hence, a question that is first answered wrong and then answered correctly will have a higher rating than a question that is first answered correctly and then answered wrong. Consequently, the impact of the initial difficulty rating will fade out. Similarly, a question that has been answered wrong many times will relative quickly adapt its rating if it suddenly gets answered correctly more often, and vice versa. The higher influence of recent answers can be considered specifically important in informal settings, where the actual knowledge of players is based on random sources and not bound the any well known learning materials. Hence, changes in the knowledge sources are not under any control and it may thus be helpful if the quiz mechanism can adapt quickly to such changes.

### 2.2 Adapting the Pool Size per Level

While it is tempting to map difficulty ratings directly to difficulty levels, doing so can cause serious problems. In particular, ratings may change in a way that there is no question for a particular level. Even if there is a low number of question for a particular level, that would imply that these are shown very often. To avoid that problem, the quiz mechanism sorts all questions by difficulty rating and then divides the question set into pools, where each pool's size correlates to the number of times that level was entered.

First experiments during the design of the quiz mechanism quickly revealed that a linear mapping from the number of questions shown to the pool size for that level leads to dissatisfying results. Lower levels got too much space and thus the span in difficulty ratings for these levels was quite high. At the same time, only very few questions were assigned to the question pools on the most difficult levels. Hence, the mechanism was altered to use the square root of the number of questions shown when calculating the pool sizes. That led to a more satisfying distribution of difficulty ratings across the pools, although it also implies that questions on easier levels will be repeated more often than those on harder levels.

### 2.3 Populating an Actual Quiz Session

When a player starts a new quiz session, the full sequence of questions to be used in that session is calculated in advance. The algorithm for populating the quiz sessions is designed in a way that at most one question per category is used in one session (provided, there are enough categories available). Since the question pools on higher levels are smaller, they

may miss some categories and populating a quiz session in the desired way may become harder. Hence, the algorithm populates the quiz session in the reversed order from the hardest to the easiest question. The algorithm starts by picking a random question from the pool for the hardest level. It then continues iteratively to the easier levels and ignores all question categories that have been used so far. In the rare case that the question pool for one level only contains questions from categories that have been used already, a question is picked randomly. As mentioned above, an implication of that mechanism is that there should be at least as much different categories available as their are levels in the quiz game.

### 3 QUIZ PRESENTATION AND IMPLEMENTATION

An implementation of the quiz mechanism has been created for the website of a living history society from Germany. The society primarily works together with museums and schools to create impressions of ancient Roman life in the first century AD for a general public audience. To support these activities it has a website that provides some texts with factual knowledge on various topics of ancient Roman history.

A first implementation of the quiz has been added to that website in 2002. It has been updated several times to keep up with general changes in the layout and structure of the website. The latest re-implementation was deployed in September 2019 (see Fig. 1 for a sample screenshot). Through all years, the same database has been used that accumulated the data from more than 100'000 quiz sessions that way.

The quiz page provides a textual introduction with instructions on how to play the quiz. Website visitors need to click explicitly on a start button to see the first question, which gives a quite clear definition on when a quiz session starts. The quiz presents its questions in 15 levels and uses multiple-choice questions with four answer options. Three times per session players may use a joker that removes two wrong answer options, thus giving them a 50% chance to answer correctly by pure guessing. The jokers are thus more generous than the prune strategy used in GAM-WATA, where only one of four options is removed (Wang, 2008). Notably, the quiz mechanism makes no restrictions on the type of jokers used.

Each quiz session ends immediately after a wrong answer with an according textual message. Levels 5 and 10 are marked as milestones and a graphical award is included in the textual message once a player has reached one of these milestones. While level 5 is

associated with a small award, level 10 is associated with a medium award. If players answer correctly in the final question on level 15, they get a large award.

The quiz currently contains 596 questions that are distributed over 25 categories. Each category is associated with a particular topic and contains at least ten questions. Most topics reflect a specific aspect of ancient Roman history, but there is also a topic about the society that hosts the quiz to motivate visitors to browse through the website. More details on the quiz content will be discussed in section 4.2 below.

## 4 USAGE DATA AND EXPERIENCE

The fact that the quiz has seen more than 100'000 quiz sessions allows for a detailed inspection of data to analyze the performance of the quiz mechanism.

### 4.1 General Usage Data

Since the adaptive mechanism is based on the number of questions played for each level, these numbers are available directly from the quiz database. Figure 2 shows the number of questions shown for each level as of February 14th, 2022. 105'576 questions for level 1 have been shown, which is equal to the number of quiz sessions. The number of questions shown decreases monotonously for each level to reach 6630 on level 15. Hence, players see the final level in only about 6.3% of all sessions.

Due to the changes in the implementation in 2019, more global usage data can be found in the server log files from that point on. In particular, the number of visits for the quiz page, the number of quiz starts, the number of answers and the number of jokers used is recorded separately. Table 1 provides a summary of these figures for the full year 2021. Data has been taken from the aggregated server log summaries.

Website visitors visit the quiz page on average in about 2.8% of all website visits. The only remarkable exception could be observed in January and February 2021, when the quiz page was visited in about 4.9% and 4.7% of all website visits. The likely reason is that the fact that the quiz had been played 100'000 times had been announced on the society's Facebook page in January 2021 and probably attracted some visitors.

There is a large variance in how often the quiz is actually started. The average value is 2.1 quiz starts per quiz page visit, but the actual values range between less than 0.7 in April 2021 and more than 5.0



Figure 1: Screenshot from the website presenting the quiz (<https://www.roemercohorde.de/de/quiz/>). The quiz page provides a textual introduction on how to use the quiz. Below the text, it presents the current level indicated on the bar, the current question, four answer options (A - D), and two buttons to use a joker and restart the quiz. Although the website also contains some English texts, the quiz is only available in German.

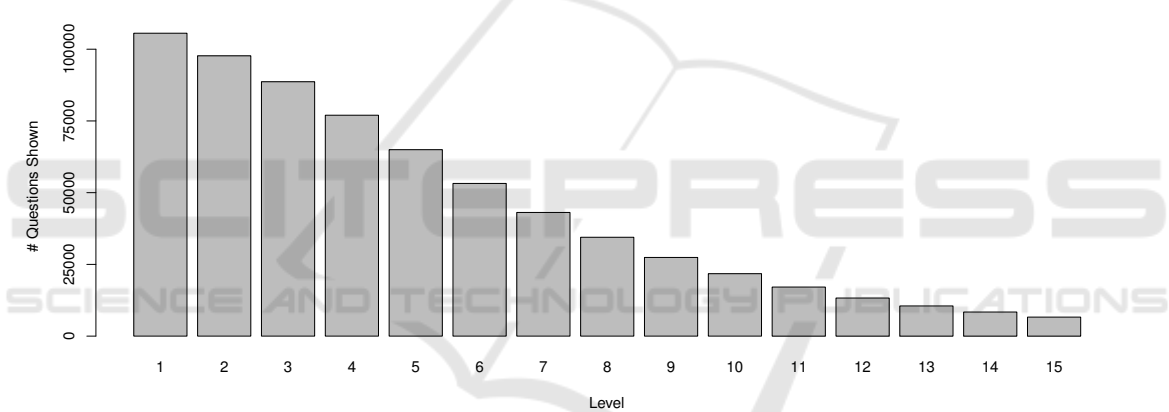


Figure 2: Number of times each level was shown in the quiz.

Table 1: Usage data for the website and the quiz in 2021 taken from the server log summaries.

Month/Year	Website Visits	Quiz Page Visits	Quiz Starts	Quiz Answers	Jokers
Jan. 2021	7939	385	885	5955	626
Feb. 2021	7055	333	354	2326	249
Mar. 2021	8597	186	313	1880	178
Apr. 2021	9309	181	123	654	68
May 2021	10065	252	586	4022	367
Jun. 2021	9949	246	1232	5240	509
Jul. 2021	8598	160	339	2574	110
Aug. 2021	8668	178	443	3177	323
Sept. 2021	8046	236	663	1647	220
Oct. 2021	7154	172	351	2688	282
Nov. 2021	7654	197	409	2763	302
Dec. 2021	7035	178	144	1078	87

Table 2: Number of different questions associated to each level and rating data for the levels.

Lvl.	No. of Questions	Avg. Rating	Min. Rating	Max. Rating
1	66	0.9523	0.9336	0.9976
2	64	0.9193	0.9054	0.9335
3	60	0.8936	0.8818	0.9052
4	57	0.8718	0.8627	0.8817
5	52	0.8515	0.8390	0.8626
6	47	0.8250	0.8138	0.8387
7	42	0.8013	0.7891	0.8134
8	38	0.7781	0.7662	0.7890
9	34	0.7579	0.7509	0.7645
10	30	0.7423	0.7354	0.7508
11	26	0.7271	0.7189	0.7349
12	24	0.7144	0.7068	0.7188
13	21	0.7017	0.6976	0.7068
14	18	0.6948	0.6921	0.6976
15	17	0.6442	0.4285	0.6918

quiz starts per quiz page visit in June 2021. The average length of a quiz session is about 6.2 answers and the players use about 0.6 jokers per session on average. More detailed usage figures are examined in section 4.3 below.

## 4.2 Content Data

As mentioned above, the quiz currently contains 596 questions that are distributed over 25 categories. The question set has not seen much changes in recent years and only few questions have been added to the question set. Hence, there are only 6 questions that have been answered less than 200 times. On average, each question has been answered about 1107 times. The most frequently used question has been answered 2506 times.

Table 2 shows the available questions per level as well as the minimum, maximum and average rating per level as of February 14th, 2022. Since the quiz mechanism calculates the available questions per level directly from the number of actually shown questions per level, numbers decrease monotonously similar to figure 2, but softened due to the use of the square root function.

The first and the last level show the largest differences between minimum and maximum rating per level with values of 0.064 for the first level and 0.26 for the last level. On most levels, the difference is between 0.01 and 0.03. That implies that questions will move at most one level up or down when their difficulty rating is updated. The only exception from that is level 14, where the difference between mini-

um and maximum is only 0.0055. A question from level 13 may thus drop to level 15 directly if answered wrong. Likewise, a question from level 15 may jump to level 13 if answered correctly.

Besides that, some ideas for improvements can be derived from these figures. For example, the average rating per level decreases by about 0.02 in most cases, except for levels 11 to 14, where the differences are close to 0.01. This indicated that there are relatively much questions that fit on that level, while there might be a need for some more easy questions for the lower levels. Adding such questions will also prevent questions from moving more than one level, because level 14 will then cover a wider range of ratings.

Table 3 shows the 25 topics of the questions, the number of questions in each pool and their minimum, maximum and average rating. The ratings show that there are both easy and hard questions in all pools, where the minimum rating is 0.78 or lower in all cases and the maximum rating is 0.92 or higher in all cases. The easiest topic seems to be “Famous Quotes” with an average rating of 0.88, while the hardest topic seems to be “Art and Literature” with an average rating of 0.76. The latter also contains the hardest question of the quiz with a rating of 0.43: “Who wrote the *Dialogus de oratoribus*?” - (A) Sueton, (B) Ovid, (C) Pliny the Elder, (D) Tacitus. Notably, that is the only question with a rating below 0.5.

While data shows that the quiz pool seems to be quite balanced in general, it can again be used to derive concrete hints for improvements. For example, the average rating of “Art and Literature” (0.76) is below the minimum rating of “Famous Quotes” (0.78). This suggests to add some easy questions to the former one or some harder questions to the latter one.

## 4.3 Quiz Performance Data

The analysis so far demonstrates that the content seems to be arranged as expected. A closer inspection of the usage data is necessary to find out whether the behaviour of the quiz players is indeed as expected. For that purpose, all 5842 quiz sessions played during the year 2021 have been extracted from the server log files for a detailed analysis.

The bar plot in figure 3 summarizes the length of each session in terms of given answers. The x-axis includes a bar for 0, because visitors may end a session without giving any answer at all. Since the log files do not reveal whether an answer was correct or wrong, it cannot be distinguished where a session ends due to a wrong answer or because a visitor stopped playing or just started a new session. The notable size of the bar for level 15 is due to the fact that the quiz ends after

Table 3: Number of questions and rating data for the 25 question categories used by the quiz.

Topic	# Questions	Average Rating	Minimum Rating	Maximum Rating
Military Equipment	30	0.85	0.70	0.97
Geography	41	0.81	0.65	0.94
Letters and Numbers	12	0.81	0.70	0.97
Politics	23	0.83	0.70	0.96
LEGIO VI VICTRIX	12	0.79	0.71	0.92
Famous Quotes	23	0.88	0.78	0.94
About the Society	11	0.85	0.74	0.92
Abbreviations	10	0.84	0.73	0.94
Roman Provinces	19	0.79	0.69	0.93
Religion	24	0.85	0.64	0.98
Craft and Engineering	22	0.84	0.68	0.97
Military Organization	31	0.81	0.62	0.95
Buildings and Architecture	24	0.87	0.69	0.98
Every-day Life	42	0.87	0.69	1.00
Drinks and Beverages	18	0.86	0.71	0.97
Wars and Battles	24	0.86	0.70	0.95
Famous Persons	39	0.78	0.55	0.94
Calendar and Time	13	0.82	0.70	0.96
Art and Literature	18	0.76	0.43	0.97
Romans Today	22	0.84	0.73	0.98
Cultures and Nations	15	0.86	0.72	0.93
Money and Economy	17	0.80	0.64	0.95
Miscellaneous	12	0.87	0.69	0.96
Latin Vocabulary (Military)	49	0.81	0.68	0.93
Latin Vocabulary (Civilian)	46	0.81	0.70	0.94

15 levels in any case, independent of the correctness of the given answer.

Section 4.1 already mentioned that the average session contains 6.2 answers according to global usage data. This corresponds to the bar plot, where the largest bar denotes sessions that end after six answers. The form of the bar plot roughly follows a Gaussian distribution and corresponds to the slope of the bars in figure 2, where there is a stronger decrease per level from levels 3 to 10 and a lower decrease on the earlier and later levels.

The box plot in figure 4 provides median values and quartiles for the time between showing a question and getting an answer. The y-axis is cut-off at 80 seconds and thus the plot does not include larger outliers. The server log files capture timestamps for events in the granularity of seconds and thus there are often the same values for two or more adjacent levels. Nevertheless, the plot shows a general trend of steadily increasing times. Most players seem to spend less than 20 seconds for an easy question and less than 70 seconds for a difficult one. Many players even answer the easy questions in about ten seconds and the harder ones in 30 seconds.

The bar plot in figure 5 depicts the probability that a player uses a joker on a given level. Probability is calculated by the number of jokers used on a particular level through the number of sessions that reached at least that level. There is a clear increase in the probability with increasing level, indicating that players tend to use jokers primarily on the difficult levels.

#### 4.4 Discussion

Content data shows that a total amount of slightly less than 600 question is sufficient to create 25 balanced question pools. Hence, the quiz mechanism has no problems in constructing a series of 15 question of increasing difficulty for each session while using at most one question per topic. Data shows that the quiz mechanism has indeed managed to organize questions into levels as intended. There is a fair increase in the difficulty rating from one level to another and there are no obvious gaps that cause an abrupt change in difficulty from one level to another. Moreover, the data can be used to derive useful hints on how to improve the question set.

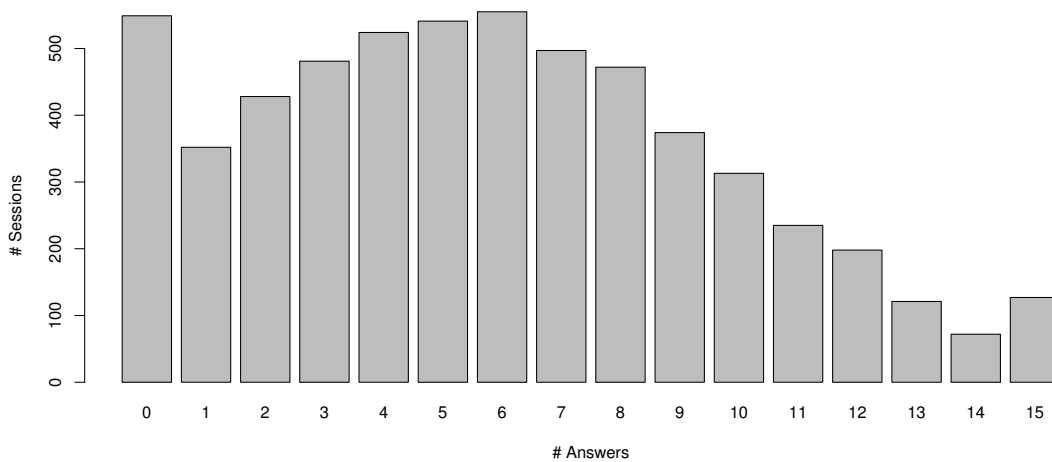


Figure 3: Number of quiz sessions played in 2021 that have terminated after the given number of answers. Numbers include sessions that stopped with a wrong answer, a correct answer on level 15 or no answer at all.

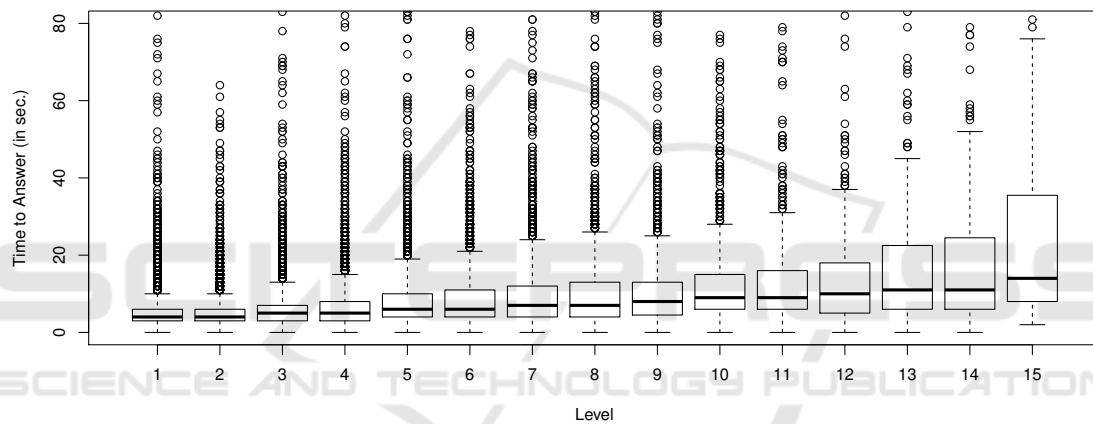


Figure 4: Median and quartile of time between showing a question and getting an answer (in seconds), based on data recorded for the full year 2021.

Performance data shows that the quiz contents are not only organized as intended, but that also the behavior of the quiz visitors is as expected. The times for an answer basically increase from one level to another. At the same time, median values for these times are quite small and thus visitors are likely kept in a smooth flow of playing one question after another also for the more difficult levels. Similar to the increase in times, also the probability to use a joker increases from one level to another. Since using a joker is an explicit action, data supports the idea that users also recognize the increase in difficulty. At the same time, the average amount of about 0.6 jokers per sessions indicate that players refrain from using jokers too often and are motivated to find (or guess) the correct answer on their own. Notably, the increased usage of jokers may be caused by other factors as well. For example, players may be afraid to loose streak and thus tend to use jokers primarily on higher levels. How-

ever, this would particularly make it attractive to used jokers to reach the milestones at level 5 and 10, but this is apparently not the case.

The overall impression, that the quiz is indeed as motivating and entertaining as expected is also supported by the fact that only about 10% of all session end after less than three answers and a fair amount of sessions manages to pass successful through about a half of the levels.

## 5 CONCLUSION

The paper presented and discussed a self-adaptive mechanism for serious quiz games. Different to related work, the algorithms does not maintain user models for individual players and does not try to adapt quiz difficulty to these models. Instead, it adapts the characteristics of the item pool and individual items

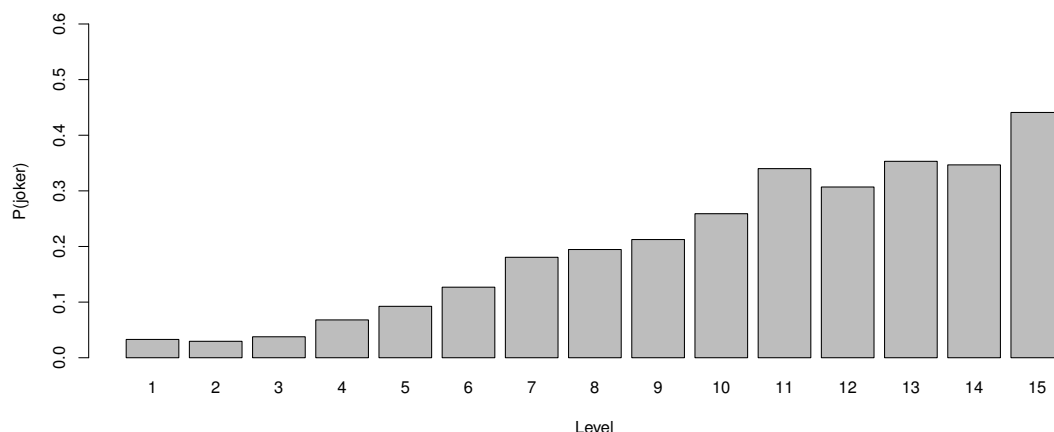


Figure 5: Probability that a player uses a joker on a given level, based on data recorded for the full year 2021.

to optimize the quiz flow for all players. Empirical data demonstrates that the mechanism is indeed able to adjust difficulty ratings without manual intervention. Similarly, the mechanism is able to populate quiz sessions in a way that the players seem to experience a motivating increase in difficulty from one level to another. This proves that a quiz can be run that way with very little resources for constant maintaining. At the same time, data recorded by the quiz mechanism can also be used to gather hints on what kind of questions should be added to improve the quiz.

Since the actual quiz implementation is used in an informal setting, there is no possibility to measure an explicit learning gain among the players. However, there is also no reason to assume that the general benefits for serious quiz games should not apply to that specific instance. A closer inspection of the players motivation and attitude towards the quiz game by using a survey associated with the quiz game is subject to future research.

## REFERENCES

- Becker-Blease, K. A. and Bostwick, K. C. P. (2016). Adaptive quizzing in introductory psychology: Evidence of limited effectiveness. *Scholarship of Teaching and Learning in Psychology*, 2(1):75–86.
- Heitmann, S., Obergassel, N., Fries, S., Grund, A., Berthold, K., and Roelle, J. (2021). Adaptive practice quizzing in a university lecture: A pre-registered field experiment. *Journal of Applied Research in Memory and Cognition*, 10(4):603–620.
- Lin, F. (2020). Adaptive quiz generation using thompson sampling. In *Third Workshop eliciting Adaptive Sequences for Learning (WASL 2020)*.
- Mangaroska, K., Vesin, B., and Giannakos, M. (2019). Elo-Rating Method: Towards Adaptive Assessment in E-Learning. In *IEEE 19th International Conference*

*on Advanced Learning Technologies (ICALT)*, pages 380–382.

- Park, J. Y., Joo, S.-H., Cornillie, F., van der Maas, H. L. J., and den Noortgate, W. V. (2019). An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments. *Behavior Research Methods*, 51:895–909.

- Simon-Campbell, E. and Phelan, J. (2016). Effectiveness of an Adaptive Quizzing System as an Institutional-Wide Strategy to Improve Student Learning and Retention. *Nurse Educator*, 41(5):246–251.

- Wang, A. I. and Tahir, R. (2020). The effect of using Kahoot! for learning – A literature review. *Computers & Education*, 149:103818.

- Wang, T.-H. (2008). Web-based quiz-game-like formative assessment: Development and evaluation. *Computers & Education*, 51(3):1247–1263.

- Wickline, V. B. and Spektor, V. G. (2011). Practice (Rather Than Graded) Quizzes, With Answers, May Increase Introductory Psychology Exam Performance. *Teaching of Psychology*, 38(2):98–101.