Examining Gender Differences in Game-Based Learning Through BKT Parameter Estimation

Saman Rizvi, Andrea Gauthier, Mutlu Cukurova, and Manolis Mavrikis

UCL Knowledge Lab, University College London[1], London, UK
{saman.rizvi, m.mavrikis}@ucl.ac.uk

Abstract. The increased adoption of digital game-based learning (DGBL) requires having a deeper understanding of learners' interaction within the games. Although games log data analysis can generate meaningful insights, there is a lack of efficient methods for looking both into learning as a dynamic process and how the game- and domain-specific aspects relate to contextual or demographic differences. In this paper, employing student modelling methods associated with Bayesian Knowledge Tracing (BKT), we analysed data logs from Navigo, a collection of language games designed to support primary school children in developing their reading skills. Our results offer empirical evidence on how contextual differences can be evaluated from game log data. We conclude the paper with a discussion of design and pedagogical implications of the results presented.

Keywords: Game-based learning, Gender differences, BKT.

1 Introduction

There is a growing interest in digital educational games as an emerging pedagogy often referred to as digital game-based learning (DGBL). While data from DGBL environments have been used for automating sequencing and feedback (e.g., in adaptive systems) [1], to address research questions around how well games were aligned to educational objectives [2], research has also demonstrated the potential of using such data to derive design implications [3]. For example, [4] presents an approach relying on learning curves analysis to evaluate how the skills targeted within the game fit student performance data. Similarly, [5] explore different design variations to optimize challenge.

Concerning methodology, previous research on AIED (Artificial Intelligence in Education) algorithms showed the significant value of Bayesian methods (such as Bayesian Knowledge Tracing and variations) in terms of their effectiveness in predicting if a student has learned a specific skill by using students' logged data [6]. However, the models built for tracing students' performance might show different results for students with different demographics and individual characteristics. For instance, recently, [7] explored the equitability of knowledge tracing in relation to 'slow' or 'fast' learners. Here, we are particularly interested in applying BKT to understand the variations in student performance and engagement within different game mechanics and between different gender groups of students. The identification of potential gender differences

[1] Funded from the Strategic Investment Board of the IOE, UCL's Faculty of Education and Society. For data collection and game development funding see http://iread-project.eu

in student behaviours in DGBL environments requires a robust, practical, and reliable methodology for looking into the interaction of gender with DGBL performance and in-game behaviours for specific games. In this paper, we present an approach to analysing such differences in the context of a language learning DGBL environment with the help of pyBKT, an accessible implementation of BKT. Identification of gender bias is the first step towards its potential mitigation, and it may have important design implications in relation to specific games that tend to appeal to one gender.

2 Methodology

2.1 The Context, Data and Participants

To examine gender disparities in game-based reading development, we leveraged player demographics and data logs from *Navigo*, a collection of adaptive educational games, as part of the iRead project [8]. The iRead database comprised data collected from digital learning products in four European languages (English, Greek, Spanish and German). This study used a subset of data from the English language domain model, generated by 127 students playing *Navigo* games in ten UK primary schools. The students in our data subset belonged to three different year groups as follows: 26 students from year 1 (13 female, 13 male, age range: five to six years old), 28 students from year 2 (15 female, 13 male, age range: six to seven years old) and 73 students from year 3 (37 female, 36 male, age range: seven to eight years old). As a selection criterion, the records were selected where students' contextual information (e.g., school, class, year group, and gender) was complete. There were around equal proportion of female (n = 65, ~ 51%) and male students (n = 62, ~ 49%).

The included data were collected across five *Navigo* games, targeting two linguistic *levels:* Phonology and Word Recognition. Each linguistic level was further categorized into language *categories*: (1) Consonant Clusters, (2) Grapheme Phoneme Correspondence (GPC), (3) Syllabification from the Phonology linguistic level, and (4) Frequency and (5) Irregular GPCs from Word Recognition. Within each category, were different language *features*, the most granular unit of language. Each language feature could be exercised through different types of games, designed to promote one of three increasingly challenging linguistic *skills*: accuracy, blending, and automaticity [8]. Each of these linguistic skills required a unique approach to learning (e.g., multiple-choice questions, mix-and-match questions), consequently leading to distinct game mechanics (e.g., puzzle game, hit-the-target game) seen in the selected five games. Playing one round of any given *Navigo* game might involve several *questions*, generating multiple game *events* (e.g., start/end, correct/incorrect) (for more details see [8]).

While playing, the students covered a total of 5,027 questions across the five *Navigo* games. These students generated 3,760 unique game log entries that were further decomposable to questions, content, and game-event tables. Moreover, each game *log* recorded general information about the game *activity*, which was defined by the specific game *identifier* (game name, id), the targeted language *feature* (feature id), and the type of linguistic *skill* trained by the game mechanic, including information such as the specific question text and answer options. It also included information about the student’s performance, e.g., duration, correct/incorrect responses, and whether any in-game feedback was received. Male students were relatively more active, generating more game log entries (2,159 games covering 2,885 questions, 57.4% game log entries) vs (1,601 games covering 2,142 questions, 42.6% game log entries) for female students.

2.2 Data Analysis

The Bayesian Knowledge Tracing (BKT) algorithm functions as a Hidden Markov Model (HMM) in its traditional form and assumes a student's knowledge (often referred to as the "mastery level" in prior literature) as a binary variable showing whether or not a student has mastered a skill. The knowledge here implies a latent variable that is updated every time a student answers a problem in a learning environment, questioning their understanding of a specific skill. BKT uses four key skill-specific parameters: **p(L_0)** or p-init, also known as p(know), is the probability that the student understands the skill beforehand. **p(T)** or p-transit, also known as p(will learn), learning probability or learn rate, is the probability that the student will demonstrate skill mastery on the next opportunity. **p(S)** or p-slip, is the probability that the student will make will answer incorrectly despite having mastered the skill. **p(G)** or p-guess, is the probability that the student correctly applies an unknown skill (a lucky guess aka guessing probability).

A recent BKT variant, pyBKT [6] is a probabilistic framework where the parameters are trained (learned or estimated) using the data from students' interactions with the learning system. The framework uses the Expectation-Maximization (EM) algorithm to achieve convergence in parameter estimation. While several useful class abstractions are possible in pyBKT (for example, creating, fitting, predicting, cross-validating, and evaluating BKT models), this study used *parameter fitting* class abstraction. In the preprocessing stage, we converted the input columns from games data using the default column mapping setting. For each linguistic category, the learners started with lowest level for a particular feature (e.g., competence level = 0) and their learn rate (i.e., p-transit) was operationalized as a proxy of students' performance. We estimated guessing and slipping probabilities in five categories (referred to as skills in BKT literature) played across three game mechanics (multiple-choice, target, and puzzle). For implementation details see https://github.com/Samanzehra/iRead.

3 Results and Discussion

3.1 Learn Rate

An estimated learn rate was derived for each student. We found that, overall, female learn rate (Mean = 0.26, SD = 0.30) was higher than male students (Mean = 0.20, SD = 0.28). Figure 1(a-c) illustrates the differences in mean learn rate in the linguistic level Phonology when three categories from Phonology were learned across different games (GPC, Clusters, and Syllabification). Figure 1(d-e) shows the differences in learn rate between both genders in the linguistic level Word Recognition.

Regardless of the categories, students' learn rate remained higher in the games designed to exercise reading *accuracy*, such as those employing puzzle (e.g., Hearoglyphs) or multiple-choice (e.g., Perilous Paths) learning strategies. The learn rate was relatively lower in the dynamic games where a student was supposed to hit a target while moving (e.g., Raft Rapid Fire), which were designed to exercise automaticity in language skills. One small exception was the seldomly played category of Syllabification, a particularly challenging language category usually taught in class in advanced years groups, where the learn rate remained consistently low.

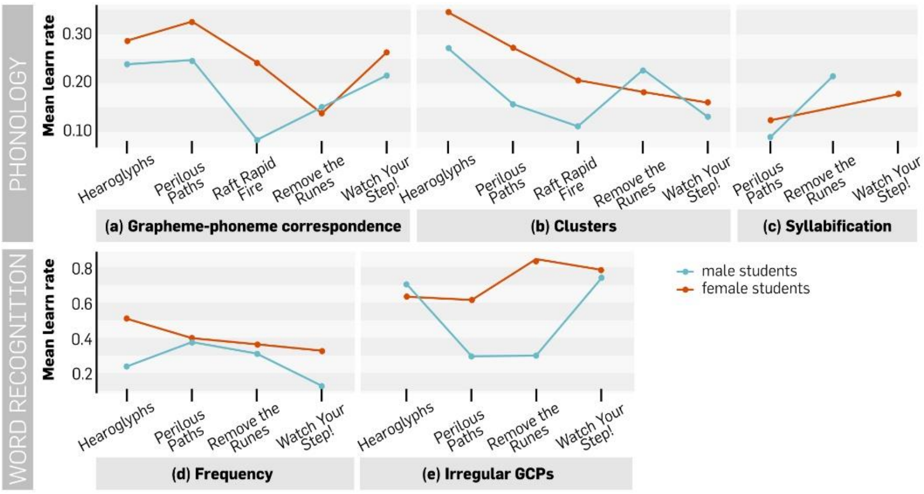

Figure 1: Gender differences in learn rate in the Phonology (a – c) and Word Recognition (d – e) linguistic levels.

3.2 Guess and Slip Rate

While learn rates were calculated for individual students, a slightly different approach was required for guessing and slipping probabilities because these probabilities may be influenced more by the type of game and its mechanics. Therefore, we evaluated guess rates and slip rates for each of the five games individually and then compared them across various categories. Figure 2 (a) to (e) illustrate the guess and slip rate in different games; each figure reports the result for one language category.

GPC was the most played language category (Figure 2 (a)), and no statistically significant gender gap was identified in the guess and slip rates. Overall, in this category, guessing probability remained higher than 0.5 except for the multiple-choice game *Perilous Paths* (where we noted p(G) = 0.43 for females, 0.48 for males). The slipping probability was also the highest for GPC in this game (p(S) = 0.33 for females, 0.38 for males). One potential reason could be that this game was the most played game by students, as it was designed for practicing skills, and in each round of *Perilous Paths*, students were supposed to cross three rope bridges (each bridge representing one question) to complete the game but perhaps due to rushing to cross the path the students made incorrect choices despite knowing the corresponding skill.

The high slipping rate remained consistent in *Perilous Paths*. However, the exact opposite trend was noticed in the puzzle game *Hearoglygh*, where, regardless of gender, students made the most guesses (p(G) = 0.77), and slipping probability remained the lowest (p(S) = 0.17). To aid focus in this puzzle game, the students were supposed to click the on-screen button to 'hear' the hidden word. Therefore, one potential reason for this contrast could be the unavailability or decreased volume of the game sounds that, particularly in classroom environments, may have been problematic. Overall, the high learn rate for the GPC category (except for the hit-the-target game; Raft Rapid Fire) may have been a result of the increased likelihood of guessing. The results indicate that just under half of the correct answers may have resulted from a guess without actual mastery of the language feature.

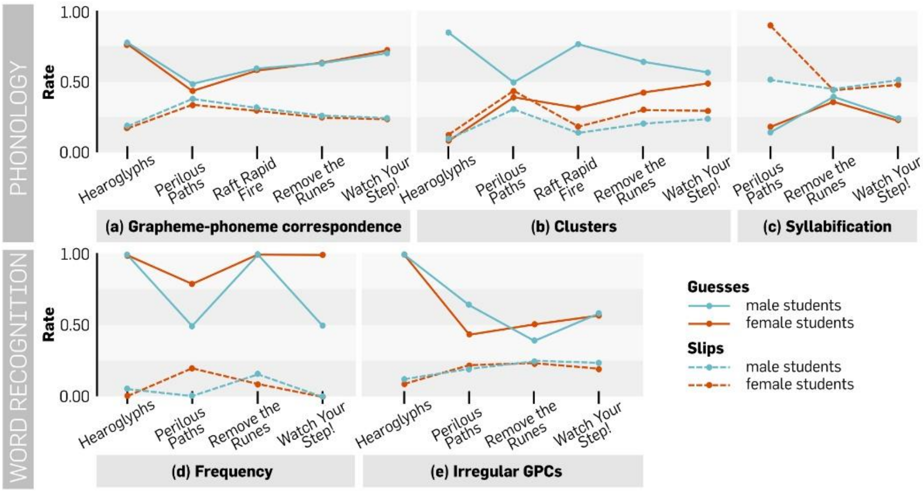

Figure 2: Gender differences in Guess and Slip rate in the Phonology (a – c) and Word Recognition (d – e) linguistic levels.

Comparing the learning behaviour when students were learning *Clusters* (Figure 2 (b)), we found identical patterns, i.e., the lowest slipping probability in the puzzle game (*Hearoglygh*) and the highest slipping probability in the multiple-choice game of *Perilous Paths* which was played most for learning new skills and practice previously learnt skills (p(S)= 0.12 for females and 0.09 for males). Also identical to the GPC category, male students made most guesses in the *Hearoglyphs* puzzle game (p(G) = 0.85), closely followed by the target game *Raft Rapid Fire* (p(G) = 0.76) when practicing Cluster-related language features. Next, the Syllabification category was played across three multiple-choice games. Unlike the rest of the two skills from the Phonology linguistic level (GPC and Clusters), students made more slips than guesses, eventually resulting in a relatively low learn rate (see Figure 1(a-c)). One potential reason could be that this category was designed to be played by relatively older students from year 3 onwards. Those students may already have some knowledge of this skill and therefore made relatively fewer guesses especially while learning in the practice game *Perilous Paths*. However, they still slipped more (for example, p(s) = 0.89 for female students, the highest slipping rate in the Phonology linguistic level).

From the Word Recognition level, the category *Frequency* was played most and across four distinct games. Consistent with the above discussed findings, students guessing probability remained highest in the puzzle game (*Hearoglyphs*). While the overall learn rate remained high for female students (Figures 1, and 2), male students made relatively more attempts to guess the correct answers in most games. Yet, the only statistically significant difference between male and female students was in the *learn rates* in the games under the *Clusters* category (H (1) = 3.844, p < 0.05). Further research is required to investigate why; it could be that, compared to other categories, "Clusters" is the least frequently explicitly taught category in English schools. This result further provides leverage for the hypothesis that the implicit learning opportunities afforded by different game mechanics may be benefitting male students more than female students (c.f. [7]).

4 Conclusion and Future Work

Like most other educational games, *Navigo* was designed with difficulty levels and game mechanics that do not necessarily favour a specific player gender on purpose. However, such differences might indeed emerge in practice [e.g., 3]. The parameter estimation methodology used in this study is a promising start in answering these and similar empirical questions in dynamic learning environments in general and DGBL in particular. The findings from this study generate data-driven insights and raise further research questions that could have been difficult to derive otherwise (e.g., through classroom observations, painstaking video analyses or other qualitative methods).

References

[1] J. C. Lester, R. D. Spain, J. P. Rowe, and B. W. Mott, “Instructional support, feedback, and coaching in game-based learning,” *Handbook of Game-Based Learning, edited by Jan L. Plass, Richard E. Mayer, and Bruce D. Homer*, pp. 209–37, 2020.

[2] E. Harpstead, B. A. Myers, and V. Aleven, “In search of learning: facilitating data analysis in educational games,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2013, pp. 79–88.

[3] X. Hou, H. A. Nguyen, J. E. Richey, and B. M. McLaren, ‘Exploring How Gender and Enjoyment Impact Learning in a Digital Learning Game’, in *Artificial Intelligence in Education*, Cham, 2020, pp. 255–268. doi: 10.1007/978-3-030-52237-7_21.

[4] E. Harpstead and V. Aleven, “Using empirical learning curve analysis to inform design in an educational game,” in *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, 2015, pp. 197–207.

[5] D. Lomas, K. Patel, J. L. Forlizzi, and K. R. Koedinger, “Optimizing challenge in an educational game using large-scale design experiments,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 89–98.

[6] A. Badrinath, F. Wang, and Z. Pardos, “pyBKT: An Accessible Python Library of Bayesian Knowledge Tracing Models,” *arXiv preprint arXiv:2105.00385*, 2021.

[7] S. Doroudi and E. Brunskill, “Fairer but not fair enough on the equitability of knowledge tracing,” in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 2019, pp. 335–339.

[8] L. Benton *et al.*, “Designing for ‘challenge’ in a large-scale adaptive literacy game for primary school children,” *British Journal of Educational Technology*, vol. 52, no. 5, pp. 1862–1880, 2021.