DOI: https://doi.org/10.1007/978-3-030-57717-9_43

Data-driven Game Design: The Case of Difficulty in Educational Games

Yoon Jeon Kim[1] and Jose A. Ruipérez-Valiente[1,2][0000−0002−2304−6365]

[1] Massachusetts Institute of Technology, Cambridge (USA) yjk7@mit.edu
[2] University of Murcia, Murcia (Spain) jruiperez@um.es

Abstract. There is increasing interest in using data to design digital games that serve the purposes of learning and assessment. One game element, *difficulty*, could benefit vastly from applying data-driven methods as it affects both players' overall enjoyment and efficiency of learning and qualities of assessment. However, how difficulty is being defined varies across the learning, assessment, and game perspectives, yet little is known about how educational difficulty can be balanced in educational games for each of the potentially conflicting goals. In this paper, we first review varying definitions of difficulty and then we discuss how we came up with a difficulty metric and used it to refine our game-based assessment *Shadowspect*. The design guidelines, metrics and lessons learned will be useful for designers of learning games and educators interested in balancing difficulty before they implement these tools in the classroom.

Keywords: Educational games · Data-driven design · Analytics.

1 Introduction

Data plays an essential role in game development and research–from qualitative user testing of early prototypes to post-launch user analytics for marketing purposes [4]. Game analytics applies computational approaches such as predictive modeling, optimization or forecasting using in-game telemetry data to provide insightful metrics [4]. As playing games has become a widely accepted activity in the broader education ecosystem [13], game analytics have been used to answer questions related to learning and assessment. Measuring how learners are learning science and math from playing the game [12, 7], and connecting in-game behaviors with other skills beyond the game environment [9]. More recently, the community calls for using game analytics to make design decisions more aligned for the purpose of learning [10]. Data-driven design that mainly relies on player telemetry data–that describes the behaviors of learners (e.g., physical location of the player in relation to meaningful objects, their interaction with others) requires a process of generating meaningful metrics.

The researchers and designers need to have clear understanding of what their design questions are and reach consensus on how to define metrics. Given the field of game analytics is fairly young [4] as is the idea of using data to inform game design [10], it remains unclear how game analytics can be used to balance among three possibly conflicting goals–learning, game, and assessment [11]. Kim and Shute [9], for example, demonstrated in an A/B testing that changing one design choice (linear puzzle sequence), significantly influences learning and enjoyment. In addition, while the game development process intends to be iterative, obtaining large enough data to be able to inform design decisions might mean less frequent iterations or too late in terms of development effort to be truly data-driven.

Difficulty is one of the core game elements that game designers carefully consider and fine tune across multiple phases of game development. For casual games, game designers are interested in knowing the difficulty of the game to adjust difficulty for the purpose of creating flow and immersed experience [7]. The game designers view that the difficulty should lie between boredom and frustration, so the players can experience a "well-shaped difficulty curve" also known as inverted U-shape curve [1], that can lead to flow state [2]. Unlike casual games, educational game designers also consider the levels of difficulty that affects not only enjoyment but learning. Gee [5] argues that players learn best in games that offer properly organized problems that push them toward the outer limits of what Gee calls their "region of competence." Hamari and colleagues similarly reported that challenge of the game has a positive effect on learning, therefore, the challenge in educational game should keep up with the learners growing abilities [6]. In assessment, difficulty is typically defined as proportion of people who answered the item correctly (also known as p-value), and existing game-based assessments mainly used p-values as the difficulty metric [9, 3]. In this short paper, the authors investigate how game analytics can be used, in the case of difficulty, to answer design questions by illustrating how metrics of difficulty were created and applied in the case of Shadowspect.

2 Case Study: Shadowspect

Shadowspect[3] is a digital 3D puzzle game where the game mechanics involve selecting a set of geometric primitives (e.g., cubes, cylinders) to recreate a 3D figure based on a number of silhouette views. The goal of the game is to allow players to build 3D figures, developing their geometric, dimensional, and spatial reasoning skills while having an enjoyable experience. The game is designed with an explicit goal of assessing math content standards and other cognitive and non-cognitive skills using the ECD framework [8]. Shadowspect collects rich telemetry data that includes all possible actions that students can do within the game environment. These data are then used to compute a number of metrics regarding students' interaction with the game. In this case study we use data from one experiment in Amazon Mechanical Turk that was aiming the design questions specified in next subsection. We collected clickstream and survey data from 32 workers from diverse backgrounds and nationalities, which had no previous experience with Shadowspect and played for around two hours be paid.

[3] https://shadowspect.org

2.1 Design Questions

The team had the following design questions related to the difficulty of the puzzles in Shadowspect: (1) What is the relative difficulty of each puzzle? (2) How can we determine the sequence of puzzles where the game has a well-shaped difficulty curve?, and (3) How long players of varying abilities would take to complete a number of puzzles? Although the team could have used game designer estimations to answer these questions, we decided to pursue a more data-driven approach in order to provide empirical and realistic estimations to teachers who want to use the game in their classrooms.

2.2 Metrics of Difficulty

To answer design questions (1) and (2), the team defined that difficulty incorporates both level of effort required by learners to solve the puzzle (i.e., more actions or time required) and relative complexity of the puzzle (i.e., fewer players can solve it). Based on this, we computed the following four metrics per each puzzle level:

- **Average time per puzzle completed**: Total time spent in puzzles divided by the number of puzzles completed correctly.
- **Average number of actions per puzzle completed**: Total number of actions performed divided by the number of puzzles completed correctly.
- **Percentage incorrect**: Number of wrong submissions divided by the total number of submissions.
- **Percentage abandoned**: Number of different puzzles that were started and not completed correctly divided by the number of puzzles started.

The four metrics have a direct relationship with difficulty (i.e., a higher value indicates more difficult), but they are in different units (e.g., percentage, time). Therefore, we generated a composite difficulty measure as follows:

1. We compute the z-scores of each metric (i.e., $z = \frac{x-\mu}{\sigma}$).
2. We sum up the four z-scores and then normalize over the maximum.

Then, the final range of our composite puzzle difficulty measure goes from 0 (easiest) to 1 (hardest). The per puzzle results of each one of these four metrics and the composite measure are available in Figure 1. The x-axis denotes the sequence in which puzzles are presented to the players. This visualization and composite measure of difficulty allowed the team to review the overall difficulty curve of the entire game and to fine tune the sequence of the puzzles. For example, the designer initially estimated that the puzzles "Warm up" and "Sugar Cones" would have a medium level difficulty, but they turned out to be much easier (as indicated by the big drop on the curve), and these two levels will be re-arranged in the next iteration of the game.

To answer design question (3), the team computed a metric including the amount of puzzles completed per unit of time, which is shown in Figure 2. This

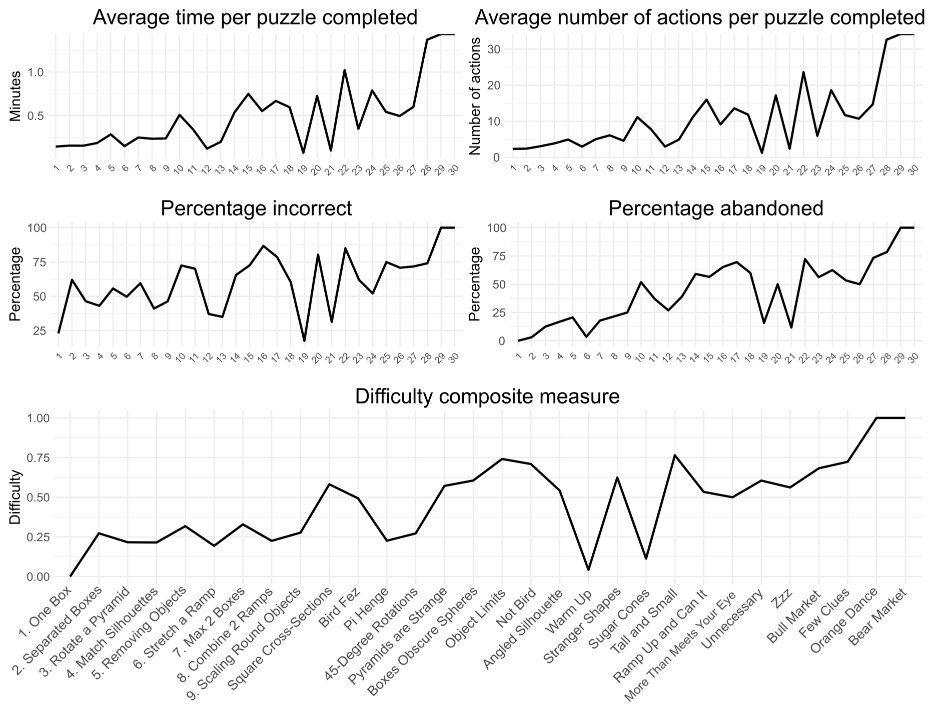

Fig. 1. Composite difficulty measure for each one the puzzles. The x-axis denotes the order in which the puzzle is presented to the user within the game.

visualization shows in a cumulative function, how many puzzles were completed with the error estimation (the sampling is performed every five minutes). This analysis allowed us to estimate how much time would be needed by users of diverse skill levels to complete an average number of puzzles. For example, based on this visualization, the team concluded that we could tell the teachers that if they perform a one hour session, their average student would be able to finish 12 ± 1 puzzles, and if they implemented a two hour session, the average student would finish 15 ± 1 puzzles.

3 Conclusions

Data-driven game design, particularly for educational games, could help designers and researchers to ensure that a game is not only playable and enjoyable, but satisfies other educational purposes such as learning and assessment. So far only little is known about this, and the field can benefit from studies that illustrate what kinds of design questions can be answered, as well as what kinds of analytics can be used and how they should be created. In this short paper, we describe how we use data-driven game design approach for the case of one game element, *difficulty*, in our game-based assessment system Shadowspect. By demonstrating how difficulty can be operationalized based on the varying notions in the literature, this work can inform future data-driven game design efforts.

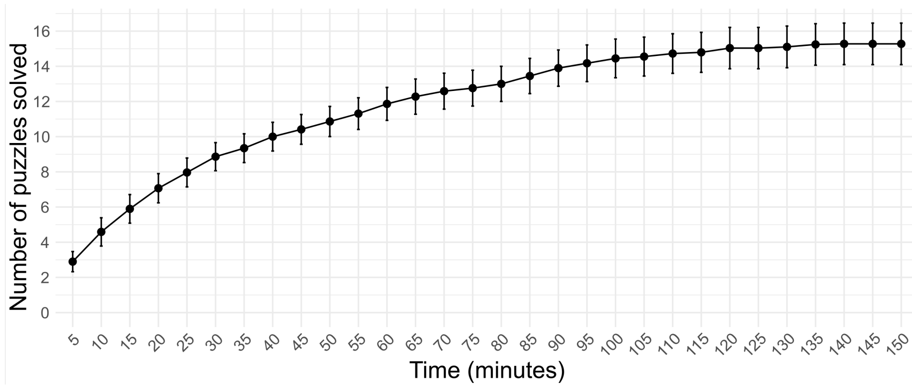

Fig. 2. Number of puzzles solved by user per unit of time with standard error deviation.

References

1. Aponte, M.V., Levieux, G., Natkin, S.: Scaling the level of difficulty in single player video games. In: International Conference on Entertainment Computing. pp. 24–35. Springer (2009)
2. Csikszentmihalyi, M.: The domain of creativity. (1990)
3. DiCerbo, K.E.: Game-based assessment of persistence. Journal of Educational Technology & Society **17**(1), 17–28 (2014)
4. Drachen, A., El-Nasr, M.S., Canossa, A.: Game analytics–the basics. In: Game analytics, pp. 13–40. Springer (2013)
5. Gee, J.P.: Learning by design: Games as learning machines. Interactive educational multimedia (8), 15–23 (2004)
6. Hamari, J., Shernoff, D.J., Rowe, E., Coller, B., Asbell-Clarke, J.: Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. Computers in human behavior **54**, 170–179 (2016)
7. Juul, J.: Fear of failing? the many meanings of difficulty in video games. The video game theory reader **2**(237-252) (2009)
8. Kim, Y.J., Ruipérez-Valiente, J.A., Philip, T., Louisa, R., Klopfer, E.: Towards a process to integrate learning analytics and evidence-centered design for game-based assessment. In: Companion Proceedings of the 9th International Learning Analytics and Knowledge Conference. pp. 204–205. ACM (2019)
9. Kim, Y.J., Shute, V.J.: The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. Computers & Education **87**, 340–356 (2015)
10. McLaren, B.M., Asbell-Clarke, J., Hammer, J.: Chi 2018 workshop: Data-driven educational game design. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–8 (2018)
11. Plass, J.L., Homer, B.D., Kinzer, C., Frye, J., Perlin, K.: Learning mechanics and assessment mechanics for games for learning. G4LI White Paper **1**, 2011 (2011)
12. Rowe, E., Asbell-Clarke, J., Baker, R.S.: Serious games analytics to measure implicit science learning. In: Serious games analytics, pp. 343–360. Springer (2015)
13. Shaffer, D.W., Squire, K.R., Halverson, R., Gee, J.P.: Video games and the future of learning. Phi delta kappan **87**(2), 105–111 (2005)