Should We Add a Progress Meter? How A/B Testing Can Support Rapid Cycles of Data-Informed Design

Derek Lomas[1,3], Dhrushit Raval[3], Steve Ritter[2], Vera van Dijk[1], Dion Dumoulin[1], Pablo Geraedts[1], Veerle Maljers[1], Isabel Mens[1], Vivek Fitkariwala[3], Nirmal Patel[3]

TU Delft[1]	Carnegie Learning, Inc[2]	Playpower Labs, Inc[3]
Delft, Netherlands	Pittsburgh, PA	Gandhinagar, India
j.d.lomas@tudelft.nl	sritter@carnegielearning.com	nirmal@playpowerlabs.com

ABSTRACT

Showing progress towards a goal is a well-established motivational design tactic. This paper describes how university students designed a new "progress meter" for an online learning game and then evaluated the effects of the design using a controlled online experiment, or A/B test.

Using the UpGrade A/B testing platform, we randomly assigned 3,200 online players to the original game or to an updated version of the game with progress meters. We hypothesized that progress meters would significantly increase student engagement, measured as the voluntary time on task (duration of play) and number of items completed. We were surprised to find that the new design significantly reduced player engagement by ~15%. Does this mean that progress meters are a sham? No. We conclude that the appropriate response to this surprising finding is to keep testing new iterations of the game mechanic. Therefore, this paper points towards a future where instructional designers and learning engineers can continuously improve online education through rapid cycles of design and A/B testing.

Author Keywords

A/B Testing; Educational Games; Gamification

CSS Concepts

• **Human-centered computing~Human computer interaction (HCI)**; User studies;

INTRODUCTION

Controlled experiments in education are typically used to evaluate new educational materials that have been developed over a period of many years (IES funding). Online experiments, or A/B tests, offer a radically different paradigm for data-informed instructional design. Namely, the opportunity to use continuous cycles of experimentation and design. A continuous cycle of designing educational solutions and testing their efficacy could lead to rapid improvements in learning software performance and new developments in the learning sciences. Online educational experiments can be used to optimize student outcomes but also used to test generalizable theories about motivation.

This paper shows an attempt to demonstrate the dual utility of A/B testing using a new open-source A/B testing platform called UpGrade. We engage university students to improve the design of an existing educational game using motivational design theory. While the game appeared to be significantly improved, the quantitative data show that students play for significantly less time with the addition of these new motivational features.

Motivational Design and Gamification

Since the early 80s, designers have incorporated video game mechanics into instructional applications [10]. Designers hope to get the same engagement and potential to motivate players in educational games. Yet, the design of successful gamification applications in education that can sustain the intended behaviour changes is still more of a guessing practice than science [4].

Background on Motivation

There are many different theories describing motivation in video games. Competition, however, is recognized as a key factor in games and gamification [14]. The satisfaction that comes with this competition makes people more likely to play a game again -- and increases self-esteem [13]. This motivation varies with age: for older children competition is more motivating than for younger children [1].

To motivate players in an educational game, simply adding winning or losing states can improve motivation [6]. While this doesn't involve other players, it still seems to address the desire for competition against one's self. Additionally, winning and losing creates more clear goals for a player.

When games have clear goals, players can be motivated by showing them progress towards those goals. A progress meter can give players an idea of how far they are in each level. It gives them a clearer goal and an idea of how long they have to play to reach the next level. This makes the players more eager to play, at least until they have reached a new level [15]. A report from O'Donovan [11] describes a variety of factors that motivate people in games. In their survey, 60% of the people were very motivated by the progress bars and 0% were very unmotivated by the progress bars.

Battleship Numberline

Battleship Numberline was created to "improve the fraction estimation accuracy of primary school children" [5]. In the

game, a player uses their number sense to estimate the location of robot pirates on a number line. Players are shown the location of the hidden enemy with a number, fraction or decimal – for instance, to find "50" on a number line from 0-100, a player would click on the middle of the line. The player must estimate where the number is on the line and try to hit it. With every hit, a coin is awarded and the hit accuracy is shown. The skull and the dots on the upper left seem to indicate the number of hits from a player, but at the moment it has no function (figure 3). The game does not show the player how many targets have been hit or how many have to be hit to complete the level. It does get more difficult when more targets have been hit. But it does not indicate at what level of difficulty a player is playing. To a player, the game is essentially endless. Figure X shows the starting menu, where you have to choose a certain topic you want to play when the game is launched for the first time. On the top of every topic it always says "completed: 100%".

Figure 1: In the original game, the player is presented with a number line marked by endpoints, e.g., 0-10. Players have to guess the location of the number presented on the console – which in the case of the first item in the game, is marked and labeled.

GAME DESIGN ITERATION

The following section describes the design of a progress bar that has the purpose of maximizing student engagement (duration of voluntary gameplay). A group of 5 first year university students in an experimental design course designed the changes presented here. Their motivation was stated as the following: "While playing the game ourselves, we have noticed that showing progress is not a big part of the game design. That is why we have decided to improve the game by showing the progress people make while playing the game. We will do this by creating an overview of levels and setting goals. The purpose of these changes is to investigate whether people will play the game for a longer period of time, and thus to improve the educational purpose of the game."

While the students left the basics of the game intact, the progress bar couldn't have been added in isolation – there had to be progress towards a specific goal. Therefore, the content of the game was divided up into a set of levels or missions that could be successfully passed or failed. This included a screen that introduced the mission goal and a screen that showed their success or failure on the mission.

Figure 2: Updated game design with functional progress meter (bottom right) and mission indicator (top left)

Change 1: Introducing "Missions".

Mission 1 is the easiest set of 10 items, while mission 2 and upwards will be progressively more difficult. The ranges on the line would change and estimating the number on the line will get more challenging. The player will have to destroy 10 ships to complete a level and will have 60 seconds to destroy a ship. Players must successfully complete each mission -- and coins earned for destroying the battleships can be seen as payment for your good work (see figure 4). A player that has 4 items wrong failed the mission.

To keep the difficulty of the items the same, within each number domain (fractions, decimals, whole numbers, etc), the items are kept the same but divided into different levels. Therefore, the differences in results before and after the change will only be caused by showing the players progress and not the difficulty of the game itself.

Change 2: Progress Meter

In the lower right corner of the screen, a progress meter shows how close a player is to completing the mission. Every time the player shoots, the target a dot lights up. Green means you have hit the target and red means you have missed. In this way, you can see how many battleships you still have to destroy to complete a mission. You can also see in which level you are at the moment on the top left of the screen.

Change 3: Mission Introduction and Ending Screens

We showed a new screen at the beginning of each mission with the objective and the time the player has for the mission (see figure 5). This gives the player a clear goal to reach during the game and gives the game a sense of purpose. At the end of every mission, whether the player has won or lost, we showed a sentence that motivates the player to go to the next mission or replay the current

mission. Below that are buttons to let the player choose either try the level again or go to the next one (see figure 11&12).

Figure 3: Mission introduction screen

Figure 4: Mission Ending Screens

RESEARCH METHOD

Experiment Design

In this study, we aimed to measure the effects of our new game designs on player engagement, which we measured as the voluntary time on task (VTOT) and the number of voluntary tasks completed (VTC). The experiment was a between-subject design – a typical online A/B test. Players were randomly assigned to the current game design (Unlimited Play) or the new game design (Missions).

The online players played the game in the same way as they normally would, choosing a topic and playing as long as they liked to. The independent variable in the research was the different game type (with or without missions and progress) and the dependent variable was the voluntary time on task and the number of tasks completed. These variables stood as a measure of the player's intrinsic motivation to participate in the game design.

Participants

The test subjects of this research were the players of the game. At the time, data told us that approximately a thousand children played the game each day. Based on that, we could assume we would have enough data to compare and make a valid conclusion, after collecting data for two days of the 'new' game. We did not ethically or legally need to gather the consent or permission from the parents to use the data as the game is educational, the research aimed at providing benefits to our users and the data is fully anonymous.. The advantage of this was that we did not have to contact a lot of participants separately. On the other hand, because of this anonymity, we could not ask them any questions about the game, their age, the environment in which they played, their motivation, and so on. Further, we could take no measure that wasn't collected from the digital experience; we couldn't gather test scores, for instance.

RESULTS

The experiment was conducted for 2 days (Friday and Monday) during which 3450 people played the game. The conditions were randomly assigned to the participants with 1673 assigned to the old game and 1777 assigned to the new game.

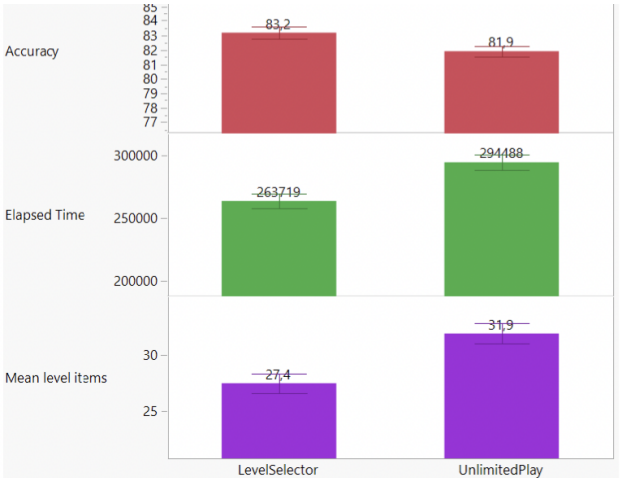

Figure 5: Bar plot showing comparison between the Level Selector version (new game design) and the Unlimited Play Version (old game design). Error bars are standard error of the mean.

The bar plot above (Figure 5) shows, from bottom to top, the number of voluntary tasks completed (Mean Level Items), the voluntary time on task (Elapsed time), and the average accuracy of player estimates. How did the new version of the game affect player motivation? Figure 5 shows that the new version of the game had an average of

27.4 items played versus the old version of the game with 31.9 items played. People who played the new version completed 4 fewer items than players in the old version without missions – a reduction in engagement of about 15%.

Elapsed Time was also significantly reduced by players in the new version of the game. On the other hand, the accuracy of player estimates in the new version of the game was significantly higher.

DISCUSSION

Data show that the average player randomly assigned to the updated version of the game played for significantly less time. This result is directly opposed to the intended effects of the new progress meter. Why might this game design element have backfired? One possible reason is that the new version of the game introduced negative feedback elements, such as the failure screen. In contrast, the open-ended play version of the game allowed for endless play without failure. When given a goal, players may stop when they reach it -- whereas, in unlimited, players may keep playing. We could test this by seeing how many items beyond 10 (the number of trials required to complete) players complete in the two groups.

Why might the progress meters have increased estimation accuracy? The content in the two conditions was the same, so it can't have been a difference in content difficulty between the two conditions (increased difficulty has been found to reduce player motivation (lomas et al). Beyond this, there are several possibilities to explore. First, it could have been a result of improved student learning in the new condition, but the effect is too strong. Second, higher accuracies may result from poorly performing players dropping out early. If remaining players have higher skill, then the accuracy will be higher overall. Third and most plausibly, the most likely reason for increased accuracy is that players are "trapped" on easier levels and therefore have to repeat the items over and over. As repetition has been found to reduce player engagement (Lomas et al), this could also account for why players didn't prefer the new mission structure.

Implications for Future Design-Experiment Iterations

The present A/B was extremely useful for showing that the updated designs were less effective as this was genuinely surprising. Based on our interpretations, we suggest a set of subsequent experiments to resolve our theoretical questions and improve the student experience.

One possibility is that the game level success criteria was simply too hard. So, instead of having all people fail the mission with 4 items incorrect, players could be randomly assigned fail with 2,4,6 or 8 items incorrect. This would plot a curve, suggesting a point for optimal level difficulty. However, the reduced motivation might not have been from the difficulty, per se, but rather the negative feedback elements in the game (red dots and "failure screen"). A new version of the game could be designed where the player can't fail missions; this would eliminate the need for the red dots or failure screen. This would maintain the repetitiveness of the game items but reduce the effects of negative feedback. It could work by simply requiring a total of 10 ships to be destroyed, no matter how long it takes. If all items in a level are deployed randomly, then design would repeat the level items until the player had a total of 10 items correct. Alternatively, players could continue on to other game missions even after failing. This would maintain the negative feedback while eliminating the repetitiveness of the game items.

Limitations

This A/B test was designed to evaluate two different game designs involving multiple design changes. Further experiments would be necessary to determine which particular elements of the different designs caused the resulting outcomes. Further, it was designed to evaluate changes in motivation not changes in learning outcomes.

This quantitative study would have been stronger with a qualitative component. Though it would have been difficult to communicate with the actual participants, we certainly could have done remote user testing sessions that involved qualitative observations of how people played. This would have allowed us to ask questions about what people liked or didn't like – or have surfaced any hidden usability issues that affected the outcomes.

One limitation of UpGrade itself was that it didn't permit direct access to the actual version of the game that was deployed. This made it difficult at times to connect the metrics we observed to the experience of the games themselves. Connecting the metrics to the experiences seems important for facilitating rapid cycles of iteration.

CONCLUSION

Designing and running online experiments are an important technique for data-informed design. To support the training of new "Learning Engineers", we sought to involve university students in the iterative design and online evaluation of player motivation in an online math learning game. In the future, it would be interesting to make the details of the experiments available to online players themselves in order to promote further STEM learning, e.g., through the communication of statistical concepts in experimental design.

ACKNOWLEDGMENTS

Thank you to Delft students for your design and research contributions. Thank you to Playpower Labs for implementing the design experiment and Arun Prakash for the visual designs. Thank you to Brainpop.com for hosting this free game.

REFERENCES

[1] Greenberg, B. S., Sherry, J., Lachlan, K., Lucas, K., & Holmstrom, A. (2010). Orientations to video games among gender and age groups. *Simulation & Gaming*, 41(2), 238-259.

[2] Bixler, B. (2006). Motivation and its relationship to the design of educational games. *NMC.* Cleveland, Ohio. Retrieved, 10(07).

[3] Chou, T.-L., & Chen, S. (2015). The effects of progress bars on diverse learning styles in web-based learning. *2015 IEEE 15th International Conference on Advanced Learning Technologies* (pp. 493-494). Hualien, Taiwan: IEEE.

[4] Dichev, C., & Dicheva, D. (2017). Gamifying education: what is known, what is believed and what remains uncertain: a critical review. *International journal of educational technology in higher education,* 14(1), 9.

[5] Lomas, D., Ching, D., Eliane, S., Sandoval, M., & Koedinger, K. (2011). Battleship Numberline: A Digital Game for Improving Estimation Accuracy on Fraction. *SREE Fall 2011 Conference Abstract* (p. 4). Washington, D.C: SREE.

[6] Lomas, J. (2014). Optimizing Motivation and Learning with large-scale game design experiments. PhD Thesis.

[7] Lomas, J. D., Koedinger, K., Patel, N., Shodan, S., Poonwala, N., & Forlizzi, J. (2017). Is Difficulty Overrated?: The Effects of Choice, Novelty and Suspense on Intrinsic Motivation in Educational Games. *CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1028– 1039.

[8] Malone, T. W. (1982). Heuristics for designing enjoyable user interfaces: Lessons from computer games. In Proceedings of the 1982 conference on Human factors in computing systems (pp. 63-68).

[9] Mekler, E. D., Brühlmann, F., Opwis, K., & Tuch, A. N. (2013). Do Points, Levels and Leaderboards Harm Intrinsic An Empirical Analysis of Common Gamification Elements. *Gamification 2013 PROCEEDINGS* (pp. 66-73). Stratford, Ontario, Canada: Center for Cognitive Psychology and Methodology University of Basel Switzerland, Dpt. of Computer Science University of Copenhagen Denmark.

[10] Mekler, E. D., Brühlmann, F., Opwis, K., & Tuch, A. N. (2013). Do Points, Levels and Leaderboards Harm Intrinsic Motivation? An Empirical Analysis of Common Gamification Elements. (pp. 66-73). Stratford, Ontario, Canada: Center for Cognitive Psychology and Methodology University of Basel Switzerland, Dpt. of Computer Science University of Copenhagen Denmark.

[11] O'Donovan, S. (2012). *Gamification of the Games Course.* Cape Town: Department of Computer Science.

[12] Olson, C. (2010). Children's Motivations for Video Game Play in the Context of Normal Development. *Sage Journals, 14*(2), 180-187.

[13] Przybylski, A. K. (2010). A Motivational Model of Video Game Engagement. Review of General Psychology. *Sage Journals, 14*(2), 154–166.

[14] Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification. using game-design elements in non-gaming contexts. *In CHI'11 extended abstracts on human factors in computing systems* (pp. 2425-2428).

[15] Venkata, F. F.-H. (2013). *Gamification of Education using Computer Games.* Las Vegas: Springer, Berlin, Heidelberg.