Optimizing Challenge in an Educational Game Using Large-Scale Design Experiments

Derek Lomas[1], Kishan Patel[2], Jodi L Forlizzi[1], Kenneth R Koedinger[1]

[1] Human Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15212
dereklomas@cs.cmu.edu
917-544-4171

[2] Dhirubhai Ambani Institute of Information
and Communication Technology (DAIICT)
Gandhinagar, Gujarat, India 382007
kishan@playpowerlabs.com
91-972-433-2824

ABSTRACT

Online games can serve as research instruments to explore the effects of game design elements on motivation and learning. In our research, we manipulated the design of an online math game to investigate the effect of challenge on player motivation and learning. To test the "Inverted-U Hypothesis", which predicts that maximum game engagement will occur with moderate challenge, we produced two large-scale (10K and 70K subjects), multi-factor (2x3 and 2x9x8x4x25) online experiments. We found that, in almost all cases, subjects were more engaged and played longer when the game was easier, which seems to contradict the generality of the Inverted-U Hypothesis. Troublingly, we also found that the most engaging design conditions produced the slowest rates of learning. Based on our findings, we describe several design implications that may increase challenge-seeking in games, such as providing feedforward about the anticipated degree of challenge.

Author Keywords

Games; Education; Crowdsourcing; Design

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Human Factors; Design; Measurement.

INTRODUCTION

In previous periods of educational game development there were few studies regarding the effectiveness of games [6, 14]. However, contemporary games have the benefit of networked data collection and large online audiences. As a result, contemporary educational games permit large-scale online experiments investigating the effects of game design factors on both motivation and learning.

Measuring the effects of design elements (such as time limits, animations, or reward settings) on motivation can be relatively straightforward. In an experimental design with random assignment, if one version of a game results in a greater duration of voluntary play, it may be described as more intrinsically motivating or more engaging. For instance, Tom Malone's early game research [11] involved the progressive removal of design elements from games and measuring differences in the average amount of time children spent on the different game versions.

Recently, Andersen *et al.* [2] demonstrated that online games can be an excellent laboratory for investigating motivation and its measurable outcome, player engagement. They quantified player engagement as the total time spent playing a game and as the number of game levels attempted. Then, the researchers randomly assigned thousands of game players to a series of A/B tests varying different design elements. They found, for instance, that music had no significant effect on engagement, celebratory animations had some positive effect, and the presence of "bonus coins" (an optional challenge) had a strong negative effect on engagement. This last finding was surprising, as the longstanding belief was that bonus challenges would be appealing to players (and make them play longer). Their study demonstrates how online experiments can be used to empirically test common hypotheses regarding the effects of design on motivation.

CHALLENGE SEEKING IN GAMES

One interesting aspect of games is that their challenge is a motivating force [12,18,20]. In everyday life, people might be expected to minimize challenge—unless, of course, the challenge leads to greater rewards. While software design portrays "ease of use" as an essential quality, game design is based on the idea that players seek challenge.

How do games promote this challenge-seeking behavior? And why would anyone seek a challenge? In many cases, greater challenges lead to greater rewards, both tangible and intangible. For instance, one may chose a challenging activity if success in that activity leads to a greater status,(as in a sporting match). Status is a key motivational element in games and also in educational environments (e.g., scores, grades and teacher reports). Because status and challenge will often correlate, we question whether challenge alone would be motivating.

Why would someone engage in a more challenging activity, if a less challenging activity leads to a perfectly equal

CHI 2013, April 27–May 2, 2013, Paris, France.
Copyright © 2013 ACM 978-1-4503-1899-0/13/04...$15.00.

reward? Several theorists have suggested that completing challenging tasks brings greater internal rewards than completing easy tasks [7,9]. For instance, children will smile more after completing longer, more difficult word-scramble puzzles [7]. This pleasure is believed to arise from one's enhanced sense of self-efficacy or the sense of competence [4] that comes from the accomplishment.

Under this explanation, it is the *successful completion* of a challenging task that is more satisfying—not the act of *doing* something challenging. Still, a significant body of evidence indicates that, if given a choice, people will *choose* moderately difficult activities over activities that are very easy or very hard [7,17,1]. Moreover, there is evidence that people enjoy *doing* activities with moderate levels of challenge, not just completing them or choosing them. For instance, using Experience Sampling Methods to interrupt individuals' daily activities, Mihaly Csikszentmihalyi [1] found that people report high levels of enjoyment during challenging tasks; he characterized this enjoyment as a conscious phenomenon called "flow." The construct of flow and "flow states" have been widely adopted in theories of game enjoyment [20,12 ,21,18].

Recently, Abuhamdeh and Csikszentmihalyi sought to experimentally test the idea that optimal enjoyment occurs during moderately challenging activities [1]. They describe the idea in terms of an inverted-U: "…the notion that we most enjoy optimally challenging activities that are not too easy or too difficult implies a curvilinear, inverted U-shaped relation between difficulty and enjoyment, so that increases in difficulty should lead to increases in enjoyment up to an optimal level (i.e., the apex of the curve), after which further increases in difficulty lead to decreases in enjoyment."

To explore this "Inverted-U Hypothesis", Abuhamdeh and Csikszentmihalyi conducted a large-scale observational study of online chess players [1]. They quantified the challenge of each game as the numeric difference between the international chess ranking of two players. Enjoyment was measured through a survey taken by players immediately after each game. Their results showed that the greatest enjoyment occurred when players faced an opponent with a higher chess rating, but not too much higher. This provided strong evidence to support the Inverted-U Hypothesis.

Notably, the authors identified another factor governing the enjoyment of individual games: regardless of the difference in chess ranking or who won, players reported feeling greatest enjoyment while playing "close games." A closer game is one that ends with a smaller difference between the point value of the pieces taken by each player. Therefore, when the difference in points was minimal, and the outcome was most uncertain, players reported that the game was most enjoyable. This finding supports previous theory and evidence [3,18,12] about the motivating nature of uncertain outcomes.

Applying the Inverted-U Hypothesis to Online Games

The Inverted-U Hypothesis seems appropriate for predicting the effect of challenge on player motivation in videogames. Indeed, many game design guides [5,18,20] and psychological theories [1,3,4,8,17,13,20,21] suggest that optimal engagement will occur at a moderate level of difficulty. However, there is little guidance about precisely how difficult a game should be.

In our research, we use large-scale experimentation to test the Inverted-U Hypothesis in an educational game context. We seek to determine whether there is a particular degree of challenge that optimizes player engagement. To identify the optimal level of challenge, we randomly assigned players to different design configurations and modeled the effect of the varying challenge on player motivation. The Inverted-U Hypothesis predicts that engagement will be highest at a moderate level of challenge, neither too hard nor too easy (Figure 1). In these studies, we operationalized challenge as the probability of success [3,17] and motivation as the duration of voluntary game engagement [2,12].

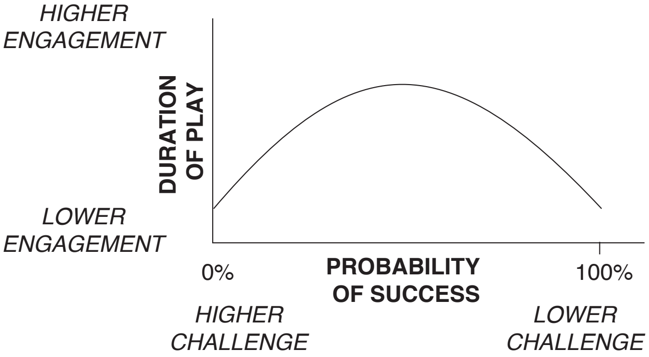

Figure 1: The "Inverted-U Hypothesis" suggests that optimal engagement (measured as duration of voluntary play) will occur at a single intermediate level of game challenge (measured as probability of success). Factors that increase or decrease challenge from this level should reduce engagement.

ONLINE EXPERIMENT 1

To document the relationship between difficulty and engagement, we used *Battleship Numberline*, an online Flash game where players attempt to explode target ships and submarines by estimating numbers on a number line. Classroom experiments have shown that the game produces significant improvements in number line estimation accuracy after only 20 minutes of game play [11].

On the basis of the Inverted-U Hypothesis, we anticipated that a manipulation of the difficulty of the game from very easy to very hard would produce an inverted U-shaped effect on engagement. To manipulate difficulty in the game, we varied two different design factors: target type and target size.

Target Type

The target type can be either a visible ship or a hidden submarine. With the visible ship (see Figure 2), a battleship is visible on the number line and players need to type a number that approximates its location. With the hidden submarine target, players are instead presented with a number indicating the location of the hidden submarine; the player then needs to click on the location of the number line that they believe corresponds to the number provided. Both the task of locating a given number on a number line and the task of naming a given location on number line are educationally relevant practice activities that may vary in their challenge.

Target Size

A larger target is easier to hit; estimates can be less accurate and still be successful. Therefore, the target size can be described in terms of "error tolerance", where a higher error tolerance indicates a larger target (Figure 2).

Take the case of a player estimating the location of a submarine that is "spotted at 20" on a number line from 0-100. If the player clicks on the number line location corresponding to 29, their estimate would have a 9% error—and an accuracy of 91%. If the target were larger, 20% of the length of the number line (an error tolerance of 10%), the estimation attempt would be successful. However, if the target were smaller (10% of the line or 5% error tolerance), the estimation attempt would miss the submarine. As a result, increasing or decreasing the size of the ship greatly affects the challenge of the game.

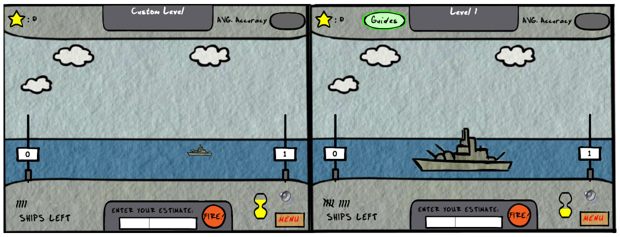

Figure 2: The left image shows the game's original ship that is 10% of the number line (5% error tolerance). The ship on the right is 40% of the line (20% error tolerance), which seemed simply too easy.

Game Design

Players are either presented with a visible ship or a number indicating the location of a hidden submarine. Once the player has typed in their estimate (in ship mode) or clicked on an estimated location (in submarine mode), a bomb falls at that location on the number line. If the bomb hits the target, there is a satisfying explosion and a gold star is released, incrementing the player's star count in the scoreboard. If the player misses, the bomb splashes in the water and corrective feedback is displayed along with the accuracy of their estimate. A running tally of the player's average accuracy is displayed in one corner of the screen and a count of the number of stars collected on the other. There is no final "winning" or "losing" state in the game—instead, players can continue to play as long as they wish. Additionally, there are no leaderboards or other mechanisms that allow players to directly compare status.

Experimental Design

To explore the effect of challenge on engagement, we constructed a 2x3 between-subjects experiment involving target size (error tolerance of 3%, 5% and 10%) and target type (ship or submarine). The experiment took place among players who had chosen to estimate whole numbers. In all conditions the players received the same 20 whole number estimation items (between 0-100) in random order; at the end of this set, the items would be repeated in another random order. Players had the option of dropping out at any time but the total number of trials was capped at 80 trials. The time limit for the ship condition was 15 seconds and 10 seconds for the sub condition.

Measuring Challenge and Engagement

Our operationalized measure of challenge was the estimated success rate of each game configuration, where success rate is measured as the percent of successful estimates divided by the total number of estimates. Engagement was measured as the average duration of play in each condition, either as the total trials played (number of estimates attempted) or the total time played (sum of reaction times across estimates).

Participants

Battleship Numberline was made available on the *GameUp* platform on Brainpop.com, which is a popular site for classroom teachers in grades 4-8. The vast majority of play occurred during school hours, with large drop offs during weekends and holidays. However, subjects were completely anonymous and were not tracked over time. As a result, the unit of experimental manipulation was on a game session, rather than a subject. A game session began when players clicked on one of the game choices (fractions, decimals or whole number estimation) and ended when players exited or made no further actions after a time-out. Each game session represented a unique experimental assignment and a single player could receive multiple versions of the game by exiting and starting over. While allowing the same individual to participate in multiple experimental assignments may appear to distort our data, the purpose of the experiment was to measure implicit user preference. Therefore, choosing to disengage from one condition and start another only enhances our measure of engagement.

Approximately 1,000 game sessions were played per day, ranging from 200 per day over school holidays to 10,000 per day when the game was "featured" by BrainPop. This scale made it possible to run a broad number of experiments simultaneously from the same subject pool. The data presented here is from one such experiment, which was conducted between 12/16/2011 to 3/8/2012 and involved 10,478 game sessions.

RESULTS

Target Size

As expected, bigger targets made the game significantly easier: the largest targets had an average success rate of 63% (sd=34%) while the smallest targets had a success rate of 29% (sd=25%). The larger and easier-to-hit targets were also significantly more engaging, resulting in 34% more estimates and 18% more time spent than the smaller and more difficult-to-hit targets (See Table 1).

While players take significantly less time to estimate the larger targets (*p*<0.001), there was no significant difference between the accuracy of their estimates (*p*=0.34). In other words, the larger targets did not appear to make players more careless in their estimates.

Targ. Size	N	Total Trials	Total Time	Accur-acy	Success Rate	React Time
3%	3462	12.6 (13.1)	60.9 (54.1)	78.8% (26.5)	29.2% (25.2)	6.8 (5.8)
5%	3479	14.6 (14.4)	66.5 (55.5)	79.5% (26.0)	43.3% (30.3)	6.6 (5.6)
10%	3537	17.0 (15.8)	71.6 (55.4)	79.6% (25.6)	62.8% (34.1)	6.3 (5.6)

Table 1. Main Effects of Target Size (with standard deviation). 3% is a smaller target than 10% and more difficult to hit. All differences are significant at *p*<0.001, except Accuracy, p=.36

Target Type

The main effects from the design factor *target type* (Table 2) are contradictory, depending on how engagement was measured: the submarine target was significantly more engaging based on the number of trials (p<0.001) but the ship target was significantly more engaging based on the amount of time (p<0.001). As the ship target asks players to type an estimate on the keyboard, each trial takes more time than the click required for the submarine target. Still, given this contradiction, it is unclear which condition can be said to be more engaging/motivating.

Targ. Type	N	Total Trials	Total Time	Accur-acy	Success Rate	React Time
Sub	5596	17.0 (16.2)	60.3 (53.8)	84.8% (15.5)	44.6% (32.1)	4.1 (2.5)
Ship	4882	12.2 (12.0)	73.3 (56.0)	73.0% (33.3)	45.9% (34.2)	9.3 (6.9)

Table 2. Main Effects of Target Type (with standard deviation). Sub involves clicking a location to estimate a number, while Ship involves typing a number to estimate a location. All differences are significant at p<0.001, except Success Rate, p=.05. The difference in N between subs and ships results from players quitting before completing one trial.

Combined Scale for Measuring Engagement

The amount of "time spent" and "challenges attempted" are useful measures of engagement [2], though they can sometimes contradict one another, as above. To resolve this contradiction, using a common psychometric practice, we created a combined measure that consists of the log transformation of the number of trials times the number of seconds of play (sum of reaction time). Though this combined measure of engagement has the limitation of making it more difficult to compare the units of engagement, it allows us to compare all design factors along a single scale. Notably, the scale balances the weight of both measures through a correlation with log(time) and log(trials) of .95 and .96, respectively. This measure corroborates that target size has a highly significant effect on engagement, but suggests that there is no significant effect of target type on engagement (p=0.57).

Challenge as a Latent Variable

To explicitly test the hypothesis that challenge has an inverted-U shaped relationship with engagement, we need a quantitative measure of challenge. Therefore, we used the average success rate of each of the six possible design configurations: the lower the observed success rate for a given level design, the greater the challenge it posed. This is similar to previous research that uses observed probability of success as a measure of challenge [3,9,17].

Figure 3 plots the relationship of challenge to engagement. In contrast to the Inverted-U hypothesis, our results show a linear relationship between difficulty and engagement (the quadratic fit was not significant, *p*=0.8). In short, the easier the game, the longer people played.

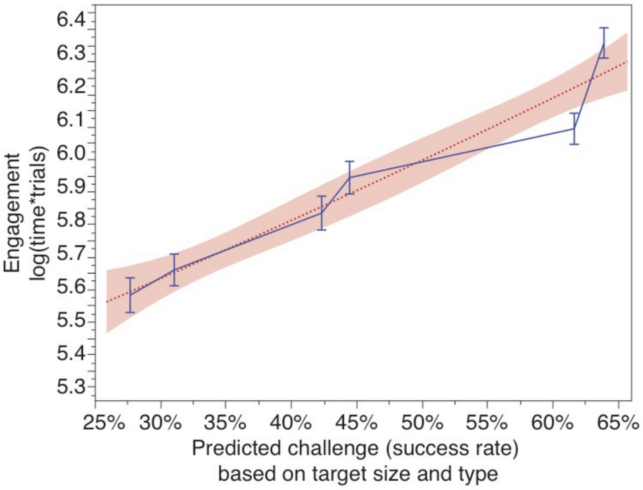

Figure 3: Relationship of challenge (predicted success rate) and engagement. Quadratic line of fit shown as dotted line with surrounding confidence of fit. The line graph plots the specific average engagement at each level of difficulty with error bars representing one standard error.

Discussion

In contrast with the Inverted-U Hypothesis, experiment 1 showed that challenge had a linear effect on engagement—the less challenging, the longer people played. However, it is possible that the game was not made easy enough, given that the average success rate of the easiest conditions were still <70%; therefore, it is possible that we only measured the left side of the Inverted U.

ONLINE EXPERIMENT 2: SUPER EXPERIMENT

To further explore the Inverted-U Hypothesis, we developed a larger experimental design that involved additional game design factors, including time limits, item sets and item sequences. Additionally, as the challenge of a game is likely relative to a player's ability (i.e., players with a higher ability will have a higher probability of success), we included an in-game pretest to formally assess player ability prior to the experiment, as a covariate.

Target Size and Target Type

We significantly expanded the range of the target size as a percentage of the number line, including the following nine sizes: 4%, 6%, 8%, 10%, 16%, 20%, 24%, 30%, and 40%. We hypothesized that the very large ship sizes would prove trivially easy to players, resulting in an inverted U-shaped curve. We maintained the two different types of targets from experiment 1 (Ship and Submarine).

Time Limit

While the time limit in Experiment 1 was consistently 10 seconds (and 15 seconds for ship targets), we tested 8 different time limits: 2, 3, 4, 5, 8, 10, 15 and 30 seconds (ship targets included an additional 5 seconds). We hypothesized that very large time limits would further reduce the challenge of the game and allow us to detect declining engagement with increased success on the right side of the Inverted-U.

Item Sets

While Experiment 1 gave all players the same set of estimation items, Experiment 2 randomly assigned players to a broad range of item sets. An item set consists of the specific target numbers that are to be estimated over the course of a level. The item sets were constructed to vary in difficulty by using data from previous experiments to create bins of items with high and low success rates [19]. Item sets also varied in the number of items presented and the endpoints of the number line. Players were randomly assigned to an item set within the estimation domain of their choice. There were 8 whole number sets, 7 decimal sets and 10 fraction sets for a total of 25 different item sets.

Item Sequencing

In Experiment 1, all items were presented in random order. In Experiment 2, we also tested three additional sequencing algorithms. Naïve0 repeats incorrect items immediately (on the next turn), while Naïve1 and Naïve2 repeats incorrect items after a delay of 1 or 2 turns, respectively.

Experimental Design

The combination of the above design factors with an ability covariate (high or low pretest score), results in a 2x2x9x8x4x25 between-subjects experiment with 28,800 possible unique configurations. Our analysis focuses on the main effects and 2-way interactions of these factors.

In-Game Pretest

After players chose to play Fractions, Decimals or Whole Numbers, they received an in-game pretest consisting of four estimation items. To construct this pretest, we used previously collected online data [19] and applied Item Response Theory methods to select items that effectively discriminated between players of different abilities and also maintained high reliability with overall performance in the game. The pretests were reasonably successful, despite having only four items, achieving a .46 correlation with the hit rate in the experiment and reliability score (Cronbach's alpha) of .62. If a player's pretest accuracy was above the median, they were labeled "high ability"; else they were labeled as "low ability."

Participants

This experimental dataset consists of game log data from 69,642 play sessions of *Battleship Numberline* that were collected from March 25, 2012 to May 4, 2012 from Brainpop's *GameUp* platform. These were players who completed the 4-item pretest described above, which occurred prior to the randomization event.

RESULTS

In an ANCOVA, all five factors (target type, target size, time limit, item sets, and item sequencing) and the pretest score had a significant main effect on the game's challenge and engagement (all *p's* < 0.0001).

Target Type

Like study 1, we found that the submarine target was significantly more engaging (p<0.001) and less challenging (p<0.001) than the ship target. Differences in engagement between target types were greatest among low-ability players (first column in Figure 4).

Target Size

Surprisingly, the extremely large targets did not lead to less engagement than moderately sized targets (second column in Figure 4). As in study 1, the larger the target was, the greater the success rate and the greater the engagement.

Time Limit

Time limit had a significant effect on player performance and engagement (third column in Figure 4). Longer time limits increased both success and engagement. The impact was greatest for the shorter time limits such that the shortest time limits were the most disengaging of any design factor.

Item Sequencing

The manipulation of item sequencing showed a significant effect on player success rates, but a minimal effect on player engagement. Naïve0 repeated unsuccessful items immediately after their first presentation, which had the effect of significantly increasing player success relative to random sequencing, for both high and low-ability players (p<0.001). Despite the improved success rate, Naïve0 had a minimal effect on engagement. Only low-ability players found Naïve0 more engaging than random presentation. (p=0.024). This result shows that improving performance may not always increase engagement, particularly when the improved performance results from repetition.

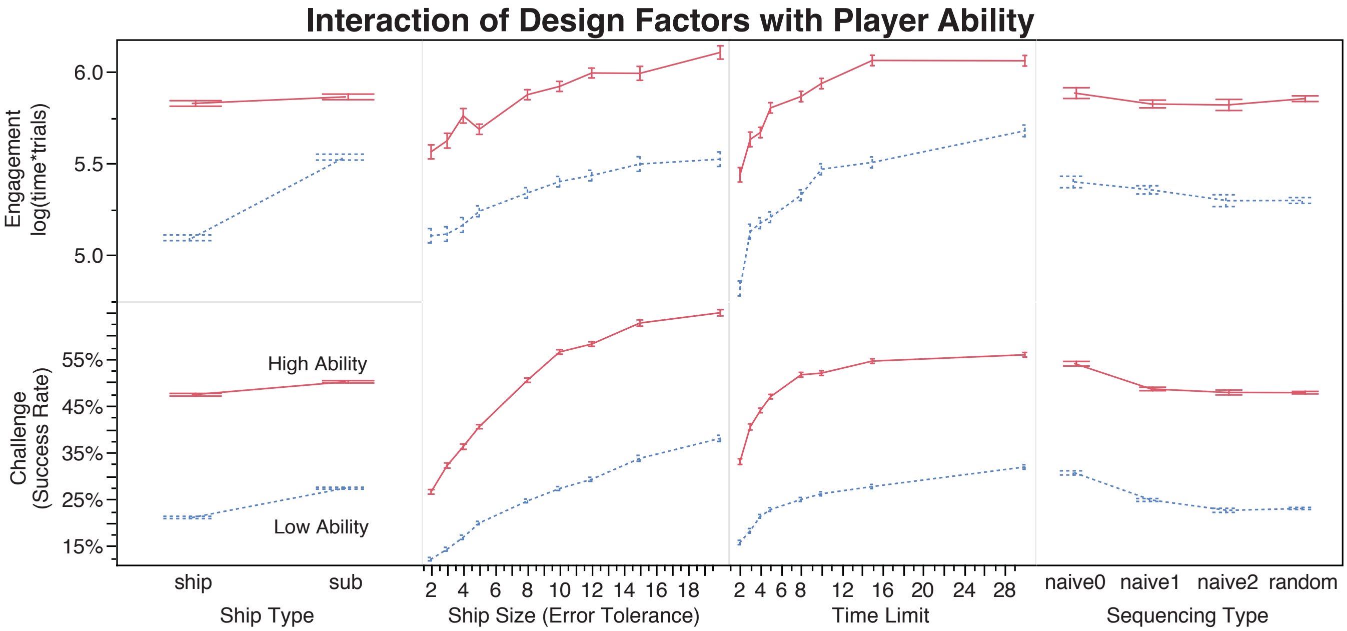

Figure 4: The effect of four design factors and player ability on challenge and engagement. The dotted lines are players with pretest scores below the median, while solid lines are players with pretest scores above the median. All four factors influence success rate and, in all cases, greater success is associated with greater engagement. Error bars show standard error of the mean.

Item Sets

As suggested by the analysis of Item Sequencing, we found that increasing the total number of items presented to a player (decreasing repetition) had the effect of improving player engagement. Additionally, we found that increasing the challenge of item sets decreases player engagement.

Item sets are the set of unique estimation items presented to players in a level. The analysis reported here focuses on a subset of fraction item sets (Figure 5). Based on previous data collection, we created an "Easy" set of 15 fractions, a "Hard" set of 15 fractions, and an "All Items" set that included easy, medium and hard items (45 total). We also deployed a set of 6 "Benchmark" fractions (1/2, 1/4, 2/4, 3/4, 1/3, 2/3). Finally, this analysis included only high ability students, as low ability students were not as sensitive to the differences between these item sets.

In the first and second columns of Figure 5, we control the number of items but vary the challenge. This shows that the easier (more successful) item set produced greater engagement. However, the third column seems to contradict this evidence, as the more difficult item set ("All Items") achieves greater engagement. This could be evidence to support the Inverted-U hypothesis, as the moderately difficult level achieved more engagement than the easy or difficult levels. However, our comparison of the third and fourth columns refutes this interpretation. Here, the difficulty (success rate) was controlled while the number of items was varied. Looking back to columns two and three, this suggests that the greater engagement of "All" results from having a greater number of items in the level.

We hypothesize that increasing the number of items increases engagement by increasing the diversity/novelty of the overall gameplay experience. This has support from at least one model of game entertainment [22], which postulates that diversity is one of the major factors accounting for the fun of games.

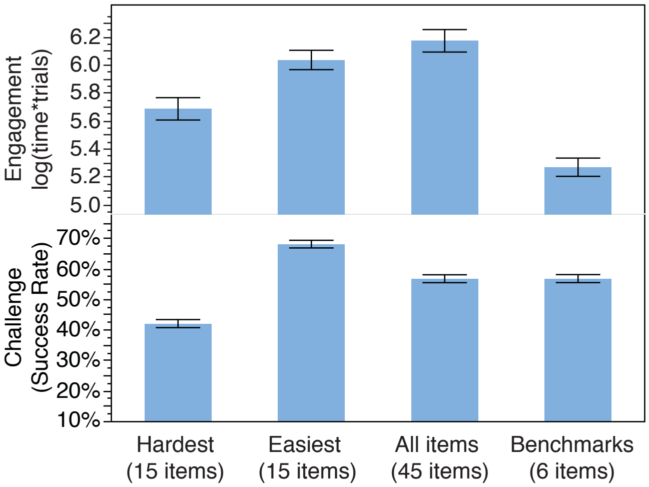

Figure 5: The relationship of challenge to engagement in item sets. When controlling the number of items (from Hardest to Easiest), easier items are more engaging. When the challenge is controlled (from All Items to Benchmarks), having more unique items is more engaging.

Challenge as a Latent Variable

Following our approach in study 1, we estimated the challenge of each level configuration, where challenge was defined as the predicted success rate of a level (lower challenge has higher success rates). The estimates produced by this model were used to plot the main effect of challenge, as a latent variable, on overall player engagement (Figure 6).

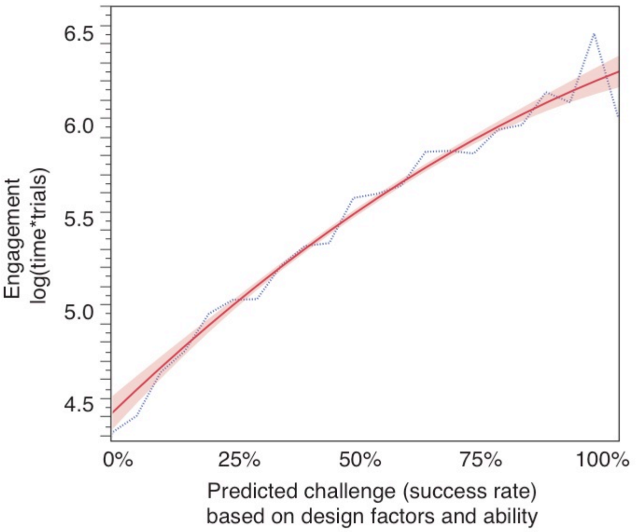

Figure 6: Plotting the effects of challenge (predicted success rate) on player engagement. Graph shows quadratic line of fit and a smoothed mean (dotted line). This shows that as the predicted success rate increases (and challenge decreases), players are likely to play for longer. In contrast to the Inverted-U Hypothesis, low levels of challenge (high success rates) never appear to negatively affect player engagement.

Success rate was predicted based on a multiple linear regression model, which included main effects of all our design factors (target size, target type, time limit, item sets and item sequencing) and the 2-way interactions between all the above design factors. To improve validity, this model was weighted by the number of trials played by each player. This produced a model with R^2=.34. As a player's ability (high/low) was expected to impact the predicted success rate of a level, we expanded our model by adding this factor and its interactions with the design factors. This improved the fit substantially, R^2=.46.

This model of challenge (predicted success rate), which involved all design factors and player ability, was used to produce the x-axis values in Figure 6. This graph illustrates the estimated engagement across the range of estimated challenge. As such, it shows that players are motivated to play for longer as the game gets easier. While the effect of game challenge is slightly curvilinear, it is not U-shaped – the model does not predict a point where low challenge reduces player motivation. Taken together, this evidence suggests that we may want to reject the generality of the Inverted-U Hypothesis.

Analysis of Learning Curves

What's wrong with making the game very easy? One hypothesis is that low levels of challenge could reduce the rate of learning. To address this question, we operationally defined learning as improvement over time, or specifically, improvement in estimation error over the log of practice opportunities (*log of opportunities* is typically the x-axis in plots of learning curves, [15]), where error was measured as the absolute value of the estimated number minus the actual number, divided by the length of the number line [16].

To address the question of whether players with large targets learned at a different rate than players with small targets, we plotted and compared their learning curves. To do so, we analyzed a subset of data that tracked each player's entire sequence of estimation attempts. This subset consisted of 1392 players who were assigned to random sequencing and who played over 30 trials in the decimal domain. We only analyzed learning in their first 30 trials to prevent player attrition from significantly affecting the measurement of learning curves. To achieve greater power, we binned the 9 different target sizes into large, medium and small targets.

Our analysis shows that larger targets result in a slower rate of learning, as compared to smaller targets; i.e.. target size has a significant interaction with the rate of learning (p=0.02). Therefore, while the largest targets were optimal for engagement, they do not appear to be optimal for learning.

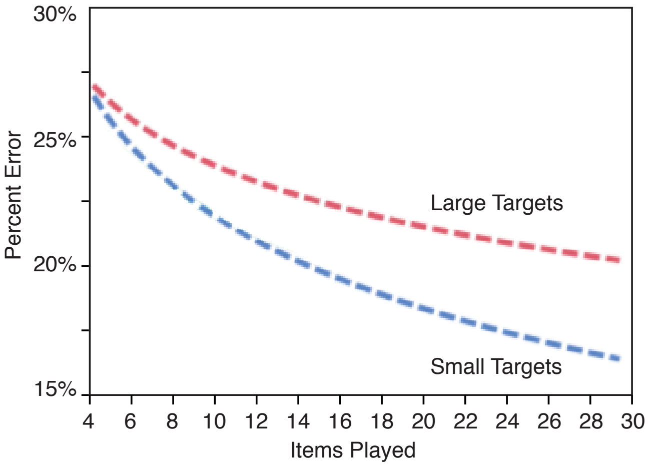

Figure 7: Learning curves show players' learning (reduction in error) over 30 items. The model suggests that larger and easier targets produce a slower rate of learning in comparison to the smaller and more challenging targets.

Limitations

There are several limitations to our approach. For instance, one could argue that there is a different *kind* of learning occurring when estimating large and small targets, making this a comparison of apples and oranges. Secondly, our claims are only meaningful for players who played for >30 trials—and this may represent a particular subset of the population that does not generalize more broadly. Finally, the high rates of attrition in online experiments make this comparison subject to critique. For example, we found that low ability students had a significantly steeper learning curve (p=0.002) than high ability students; if there were a greater proportion of low ability students in the small and medium conditions, this might suggest that the effect occurred as result of selective attrition. In this case, however, the proportion of low-ability students differed across conditions by only 2 percentage points—a difference too small to account for our reported effect. The above critiques notwithstanding, our preliminary evidence suggests that the easiest targets, which were optimal for engagement, were not optimal for learning.

DISCUSSION

In two experiments, we systematically manipulated the design space of an online learning game to determine the optimal level of challenge for supporting maximum engagement. In contrast to the Inverted-U hypothesis, which predicts that a moderate level of challenge should lead to maximum engagement, we found that the easier the game, the longer people played. This is a surprising finding, given the substantial amount of theory [4,9,8,11,20] and evidence [1,3,7,17] that suggests that moderate levels of challenge will lead to greater motivation.

It is possible that, despite our efforts, we never made the game easy enough. The Inverted-U hypothesis could still stand if the theoretically optimal success rate is very high (>90%)—in that case, we may not have observed the negative effects of decreasing challenge because we lacked sufficient data in this area of very high performance. In many video games, for instance, the vast majority of user actions are rewarded, with only an occasional setback.

Still, past researchers predicted and observed much lower optimal success rates. For instance, Atkinson [3] predicted that motivation would be greatest when the uncertainty of success is highest (i.e., a 50% probability of success). Shapira [16] predicted that the optimal success rate would be even more difficult, from 25-40% probability success depending upon individual differences. Furthermore, Csikszentmihalyi's online chess study [1] found that chess players reported greatest enjoyment when their probability of success was approximately 20% (i.e., facing opponents with a chess rating 262 points greater than their own).

Differences Between the Studies

If online chess players find a 20% success rate most enjoyable, why might our game produce greatest engagement with far less objective challenge? There are many differences between the present study and the online chess study that may account for this difference.

Random Assignment: While we randomly assigned players to a particular game configuration, participants in the online chess study freely chose their opponents. *Feedforward:* While the chess players were given information about their game's challenge (i.e., the ranking of their opponent) prior to playing, our players were not given information indicating the amount of difficulty they were encountering. *Status Opportunities*: While the chess players could improve their chess ranking by beating more advanced players, our game provided no opportunity to improve one's status. *Measure of Challenge*: Their study measured challenge using self-report and the difference of chess rankings between players. The difference in chess rankings corresponds to a probability of success (e.g., players with identical rankings have a 50% probability of success). Our study quantified challenge not as the probability of successfully winning a game, but of the probability of hitting individual targets. *Measure of Enjoyment/ Motivation*: Their study modeled the effect of challenge on self-reported enjoyment while our study modeled the effect of challenge on engagement (the duration of play). Engagement may or may not correspond to enjoyment. *Population Characteristics*: Their players were older than our players and their players likely had greater expertise in chess than our players had expertise in *Battleship Numberline*.

Theoretical Hypotheses and Design Implications

Given these broad differences between the studies, we now propose four concrete hypotheses that can add explanatory value and suggest implications for design.

Effectance Motivation Hypothesis

Our players were clearly motivated by success—the more successful they were, the more motivation they had to keep playing. The idea that success will increase motivation is predicted by the Effectance Motivation hypothesis [8]. A key implication of this hypothesis is that increasing player success rates is likely to increase player motivation. While one might accomplish this by making the targets even larger, the player might attribute their success to the game, rather than to their own achievement. Therefore, we predict that increasing success will have a greater effect on engagement if it results from a mechanism that players can attribute to their own competence. For instance, the game could improve performance by increasing in-game learning (which is, after all, the goal!). This improved learning could occur through explicit in-game instruction (e.g., a tutorial) or by providing more informative feedback (such as labeling each prior attempt, as in the constructive feedback described in Malone [12]).

Expertise Hypothesis

There is ample evidence that some groups of people demonstrate strong "challenge-seeking" behaviors (e.g., rock climbers, chess players, and video gamers). Perhaps challenge-seeking only tends to occur after individuals have acquired some significant level of expertise. Therefore, perhaps we did not observe the inverted-U because the players of our online game likely had little prior experience. This hypothesis predicts that player expertise will interact with the effect of challenge on player engagement, producing an inverted-U for players with high expertise, but not for low expertise players (who would be expected to prefer very low levels of challenge One implication for design from the Expertise Hypothesis is to make the first level of the game as easy as possible, and introducing greater challenges only after some degree of expertise has been attained.

Feedforward Hypothesis

Models of achievement motivation predict that people will attribute more value to success when the task is more challenging [1,8,9]. However, players did not have any information about the challenge of their particular game configuration (other than observations of their own performance). This is a key difference from the online chess study, where players knew their opponent's international chess ranking.

Therefore, an implication for design is to provide feedforward to players about the challenge of the task they are playing. This may allow players to more appropriately value their success and failure in the face of challenge.

Close Game Hypothesis

Abuhamdeh and Csikszentmihalyi's hypothesis [1] about "close games" may help explain our results. They hypothesize that a player's motivation will increase during close games, when there is high uncertainty about winning or losing. However, this kind of uncertainty does not occur in *Battleship Numberline*, as the game does not indicate to players whether they have won or lost (players can continue to play for as long as they like). Without winning or losing, there can be no close games. Additionally, close games only occur when the challenge of the game is closely matched to the player's ability. In other words, players with greater ability may find challenging games more engaging when the challenge increases their uncertainty about winning or losing.

The Close Game Hypothesis predicts that a clear win/lose state should make challenge motivating when the winning/losing criterion is closely matched to player's ability. As the performance criterion approaches a player's performance capacity (and when the outcome is most uncertain), the player is predicted to experience greatest motivation.

CONCLUSION

Our research investigated the effects of various game design factors on challenge, motivation and learning. While we hypothesized that moderate levels of challenge would maximize engagement (the inverted-U hypothesis) we instead observed that the easier the game, the longer people played. This is a surprising finding that prompts new research questions. For instance, given that the most engaging conditions were not the most optimal for learning, we can investigate methods for jointly optimizing learning and engagement. From a design perspective, we can investigate game design patterns that support challenge-seeking behavior, as players in more challenging conditions may benefit from a faster pace of learning.

This study contributes one approach to harnessing the crowd to optimize game designs. We used a large-scale factorial experiment and thousands of online users to identify which configuration of game design factors best supported player engagement; the same approach can be used to identify the optimal designs for supporting player learning.

Limitations

One limitation of our study was that it was susceptible to self-selection effects: it was relatively common for players to abandon games and start again, where they would be placed in a new experimental condition. This suggests that players may have deliberately exited a less preferred condition to find a more preferred condition. Though this self-selection was not intended, since the main goal of our study was to measure how design factors affected player motivation, it seems unlikely to alter our main results.

Still, online game experiments are highly limited by the fact that the data collected are from anonymous and remote subjects. This means that researchers have a very limited capacity to gain qualitative insight into *why* subjects are responding the way that they do. This suggests the importance of triangulating findings between online experiments and laboratory or field studies [20]. For instance, we may discover that "duration of play" does not, ultimately, sufficiently correlate with player enjoyment.

Our operational definition of challenge is also subject to critique. We defined the challenge of a particular level configuration as the inverse of its predicted success rate. Success rate has been used as a measure of challenge in previous research [3,17,7,8,9] and it is similar to the notion of a difference in chess ranking in [1] (since this difference corresponds to a probability of success). Still, there is evidence that *perceived challenge* is a greater predictor of enjoyment [7,1] than objective challenge. This may reflect the fact that the objective measure of challenge does not adequately capture the feeling of effortfulness that is associated with the perception of challenge.

Future Work

Perhaps increasing objective challenge (lowering probability of success) is simply not desirable to many game players. Interestingly, challenge is often correlated with increased gameplay diversity; a feature that we found to be a highly engaging. As increased gameplay diversity can outweigh the negative effects of challenge (as was the case in the "all items" level), it will be important to carefully separate these two variables in future research.

Future work should also address the validity of the operational definitions of challenge and engagement. The promise of online game research relies on valid metrics obtained in the context of anonymous gameplay. For instance, is the maximization of motivation/engagement, measured as voluntary time on task, appropriate as a data-driven game design goal? This question is particularly important for applied research goals, which seek to maximize outcome measures like engagement or learning. Future applied research may also investigate more efficient experimental designs, to minimize the cost of experimentation to designers and game players.

Very large numbers of voluntary participants gives online game research the potential to significantly expand theories of human learning and motivation. The implications for design in our discussion (in-game tutorials, feedforward about challenge, easy first levels and winning/losing states) are very standard design patterns that are widely used in games. However, by framing these game design elements as embodiments of a specific theoretical hypothesis, we can use them to test or extend theories of motivation. Therefore, it may be fruitful to systematically explore well-known game design patterns in order to generate additional hypotheses.

ACKNOWLEDGMENTS

Game Programming: Nirmal Patel and Nikhil Poonwala. Game Animation: Jeremy Galante. Statistical Support: Jonathan Steinhart. For conceptual inspiration, much thanks to Prof. Sharon Carver, Prof. Robert Siegler, Prof. Mike Cole, Prof. Mike Levine, "The Shark Game" and the playpower.org community. For advising the experimental design and analysis, much thanks to Prof. Joel Greenhouse, Prof. Trent Gaugner and Prof. Brian Junker. Thanks to Dixie Ching, Beau Dabbs, Burr Settles and Turadg Aleahmad for reading preliminary drafts. Thanks to the members of PIER and SERG for helpful feedback! Game R&D supported by funding from the National STEM Game Competition Prize, Marvell's Educational Application Grand Challenge, and the Next Generation Learning Challenge (MFDC). Graduate funding from IES Program for Interdisciplinary Education Research (PIER) and DARPA. Data analyses were conducted in the highly recommended statistical package JMP 10. Special thanks to BrainPop.com for supporting educational research!

REFERENCES

1. Abuhamdeh, S. and Csikszentmihalyi, M. The Importance of Challenge for the Enjoyment of Intrinsically Motivated, Goal-Directed Activities. *Personality and Social Psychology Bulletin 38*, 3 (2012), 317–330.
2. Andersen, E., Liu, Y., Snider, R., Szeto, R., and Popovic, Z. Placing a value on aesthetics in online casual games. *CHI 2011, May 7–12, 2011, Vancouver, Canada* (2011).
3. Atkinson, J. Towards Experimental Analysis of Human Motivation in Terms of Motives, Expectances, and incentives. In J. Atkinson, ed., *Motives in fantasy, action and society*. Van Nostrand, Princeton, NJ, (1958) 288–305.
4. Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 4,* 227–268.
5. Gee, J.P. What video games have to teach us about learning and literacy. *Computers in Entertainment*, 1 (2003), 20.
6. Hays, R. *The effectiveness of instructional games: A literature review and discussion*. 2005. Technical report 2005-004, Naval Air Warfare Center Training Systems Division, Orlando, FL.
7. Harter, S. Pleasure derived from challenge and the effects of receiving grades on children's difficulty level choices. *Child Development 49*, 3 (1978), 788–799.
8. Harter, S. Effectance Motivation Reconsidered: Towards a developmental model. *Human Development; Human Development*, (1978), 34–64.
9. Heckhausen, H. Achievement motivation and its constructs: A cognitive model. *Motivation and emotion 1*, 4 (1977), 283–329.
10. Heffernan, N. and Koedinger, K.R. A developmental model of algebra symbolization: the results of a difficulty factors assessment. *Proc. Twentieth Annual Conference of the cognitive science society*, (1998), 484-489.
11. Lomas D., Ching D., Stampfer, E., Sandoval, M., Koedinger, K. Battleship Numberline: A Digital Game for Improving Estimation Accuracy on Fraction Number Lines. *Conference of the American Education Research Association (2012)*
12. Malone, T. Toward a theory of intrinsically motivating instruction. *Cognitive Science 5*, 4 (1981), 333-369.
13. Malone, T.W. What Makes Things Fun to Learn? Heuristics for Designing Instructional Computer Games. 1980, ACM.
14. Randel, J., Morris, B., Wetzel, C., and Whitehill, B. The effectiveness of games for educational purposes: a review of recent research. *Simulation & Gaming 23*, 3 (1992), 261–276.
15. Ritter, F. and Schooler, L. The Learning Curve. *International Encyclopedia of the Social and Behavioral Sciences*, 2001, 8602–8605.
16. Rittle-Johnson, B., Siegler, R.S., and Alibali, M.W. Developing Conceptual Understanding and Procedural Skill in Mathematics: An Iterative Process. *Journal of Educational Psychology* 93, 2 (2001), 346–362.
17. Shapira, Z. Task choice and assigned goals as determinants of task motivation and performance. *Org. Behavior & Human Dec. Proc.* 44, 2 (1989), 141–165.
18. Schell, J. *The Art of Game Design.* Morgan Kaufmann (2008)
19. Stamper, J., Lomas, D., Ching, D., Ritter, S., Koedinger, K., and Steinhart, J. The Rise of the Super Experiment. *Educational Data Mining*, (2012), 196–200.
20. Sweetser, P. and Wyeth, P. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE) 3*, 3 (2005), 1–24.
21. Yannakakis, G.N., Denmark, S., and Hallam, J. Towards optimizing entertainment in computer games. (2007), 933-971