Is Difficulty Overrated? The Effects of Choice, Novelty and Suspense on Intrinsic Motivation in Educational Games

J. Derek Lomas[1,2], Ken Koedinger[2], Nirmal Patel[2],
Sharan Shodhan[2], Nikhil Poonwala[2], Jodi L. Forlizzi[2]

The Design Lab[1]
UC San Diego
dereklomas@gmail.com

HCI Institute[2]
Carnegie Mellon
{krk,forlizzi}@cs.cmu.edu

ABSTRACT

Many game designers aim to optimize difficulty to make games that are “not too hard, not too easy.” However, recent experiments have shown that even moderate difficulty can reduce player engagement. The present work investigates other design factors that may account for the purported benefits of difficulty, such as choice, novelty and suspense. These factors were manipulated in three design experiments involving over 20,000 play sessions of an online educational game.

The first experiment (n=10,472) randomly assigned some players to a particular level of difficulty but allowed other players to freely choose their difficulty. Moderately difficult levels were most motivating when self-selected; yet, when difficulty was blindly assigned, the easiest games were most motivating. The second experiment (n=5,065) randomly assigned players to differing degrees of novelty. Moderate novelty was optimal, while too much or too little novelty reduced intrinsic motivation. A final experiment (n=6,511) investigated the role of suspense in “close games”, where it was found to be beneficial. If difficulty decreases motivation while novelty and suspense increase it, then an implication for educational game designers is to make easy, interesting games that are “not too hard, not too boring.”

Author Keywords
Learning; Education; Games; Intrinsic Motivation; Challenge; Difficulty; Novelty; Suspense; Theory; Experiments; Flow; Near Win; A/B testing;

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

CHI 2017, May 06-11, 2017, Denver, CO, USA
© 2017 ACM. ISBN 978-1-4503-4655-9/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025638

INTRODUCTION

It is almost a truism that good games shouldn’t be too hard or too easy. It isn’t surprising that games can be too hard: after all, a core objective in HCI is minimizing user difficulty and maximizing ease-of-use [20,40]. That makes it much more surprising that games can be too easy. Can games really be too easy? Or just too boring?

The evidence is mixed: in some studies [4,18,1], difficulty is good for intrinsic motivation and in other studies it is not [33,44]. To clarify this issue for educational game designers and others, we ran three controlled experiments to test how various design factors modulate the role of difficulty on player intrinsic motivation.

The Inverted U-Shaped Curve Theory

To make it easier to make learning fun, Malone and Lepper [38] organized a “Taxonomy of Intrinsic Motivations for Learning”: ~30 theoretically grounded principles for designing intrinsically motivated instruction. The taxonomy’s opening claim states, “The activity should provide a continuously optimal (intermediate) level of difficulty for the learner.” This claim was based on Csikszentmihalyi’s theory of “Flow” [10], a theory that is now the basis of many contemporary theories of game enjoyment [51,48]. Formalized, Flow theory describes the relationship between difficulty and enjoyment as an inverted U-shaped curve [1].

> 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. (p. 318)

Difficulty vs Challenge

Difficulty, as a theoretical construct, has an established history; in psychometrics, it is measured as a test item’s average error rate in the population of test takers [27]. In other words, more difficult items have a greater probability of being answered incorrectly. Similarly, if a game is more difficult, one has a greater probability of losing [3]. While the quote above [1] uses difficulty and challenge interchangeably, this paper will use the term *difficulty* to mean, precisely, “the probability of task failure” and will abstain from defining *challenge*, which we believe is a more nuanced and complex concept.

Observing the Inverted-U Theory in Online Chess

To test the theory that difficulty has an inverted-U effect on enjoyment, Abuhamdeh and Csikszentmihalyi [1] conducted an empirical investigation of thousands of online chess players. They measured difficulty as the difference between the ELO chess ranking of two players (two players with the same ELO score have a 50% chance of winning). Enjoyment of games was measured via self-report immediately after each game. The result? Players most enjoyed games when they had an 80% chance of losing. These players liked hard games—but not too hard.

A few factors cloud these results. First, it is noteworthy that players had an incentive to play against harder players: winning against a higher-ranked player increases one's own ELO chess rank. Secondly, players were not blindly or randomly assigned to their opponents; instead, players could choose their opponents based on their chess rank. This gave players the ability to control their expected game difficulty. The ranking incentives and awareness of difficulty may have affected the results if certain types of players self-selected more difficult opponents. Would the results be different if ran as a controlled experiment?

Testing the Inverted-U with Random Assignment

Several years ago, we sought to replicate this inverted-U effect using a controlled experiment in the educational game "Battleship Numberline" [33]. We hoped to identify the optimal level of game difficulty by randomly assigning hundreds of different game versions to >50,000 online players. Gameplay data revealed the difficulty (average items failed divided by total items attempted) and the intrinsic motivation (duration of voluntary play) of each game variation.

Our results were surprising, as the optimal level of difficulty seemed to be "as easy as possible." Nearly all increases in game difficulty reduced player intrinsic motivation. For both high and low ability players, easier games were consistently played longer, even when the failure rate was less than 10%. This was particularly surprising in light of previous work that indicated intrinsic motivation would be maximized when the difficulty (failure rate) of the game was between 80%-50% [1, 4].

Although our results weren't predicted by the inverted-U theory [1], they do seem to be supported by other theories of intrinsic motivation. Consider that, by definition, increased difficulty increases the rate of task failure. These failures produce negative feedback that can reduce intrinsic motivation by reducing expectations for future successes [17], reducing perceived task value [17] and reducing self-perceptions of competence [11]. Moreover, more difficult tasks are more effortful: increased effort increases fatigue and increased fatigue increases rates of task switching [31, 21]. Indeed, fatigue may be the psychological reason why harder games cause players to disengage faster [33]. In any case, these theories imply that the failure accompanying difficulty will not improve intrinsic motivation.

Still, there is plenty of evidence that challenge is enjoyable and motivating [10], particularly with success [18]. Might the positive aspects of challenge come from factors other than difficulty? For instance, when one advances through a game, the new game levels are often more difficult – but they are also new. When games introduce greater difficulty, they often also introduce interesting new design elements. Perhaps games that are "too easy" don't suffer from a lack of difficulty, but a lack of interestingness.

Research Question

Our research aims to clarify the current situation regarding difficulty optimization in games. As Flow Theory strongly predicts an inverted-U relationship between challenge and intrinsic motivation, what other factors of challenge, apart from difficulty, might produce this inverted-U shape?

In the following three experiments, we investigate three different factors associated with challenge: player choice of difficulty, novelty and suspense. In the first experiment we use the same experimental design as the chess study [1] and show that letting players self-select their difficulty can produce an inverted-U shaped curve. In a second experiment, we randomly assigned players to different degrees of novelty and find that a moderate degree of novelty maximizes motivation. This evidence supports the idea that the motivational nature of challenge may stem from the novelty found in challenge as much as from the difficulty. In a third experiment, we randomly assign players to different criteria for winning, in order to dissociate player skill from their likelihood of winning. Our results show that players are motivated by the suspense of a close game, which tends to occur when the difficulty is matched to the player's skill. In total, this paper tests six hypotheses, provided below for convenience

> **H1.1:** Providing a choice of difficulty will produce an inverted U-shaped relationship between difficulty and engagement.
>
> **H1.2:** Player choices will resemble an inverted U-shaped curve, with most players choosing moderately difficult levels.
>
> **H1.3:** Higher skilled players will self-select themselves into more difficult game levels.
>
> **H1.4:** Knowing that game levels are "very easy" will decrease player persistence relative to players who don't know the difficulty of the level.
>
> **H2:** A moderate degree of novelty will maximize player motivation.
>
> **H3:** The closer the game, the greater the suspense and the greater the intrinsic motivation.

Battleship Numberline Game Design

Battleship Numberline [33,34] is a simple online game where players attempt to explode targets by estimating numbers on a number line. In subsequent experiments, players were presented with numbers that indicating the location of a hidden submarine between two endpoints (e.g., "submarine spotted at ½" between the end points of 0-1). When a player clicks on a location along the line to indicate their estimate, a bomb falls at that location and the hidden submarine then becomes visible as feedback. 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. 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.

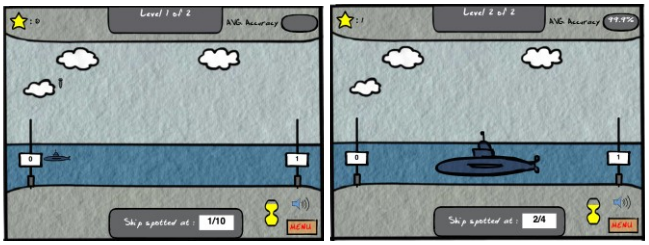


Figure 1: The *Battleship Numberline* game screen. From left to right we show the "Hard" and "Very Easy" game level. The game is easier when the target is larger because players can make more inaccurate estimates and yet still be successful.

Participants

Battleship Numberline was made available on the *GameUp* platform on Brainpop.com, a popular site for grade 4-8 classrooms. Data from experimental game sessions were collected, largely during school hours, with large drop offs during weekends and holidays (suggesting that the games were primarily played in a classroom).

The educational game site offers dozens of free games and teachers have little control over their student's activities. Thus, the decision of how long to play a particular game appears to be largely up to the student. Subjects were completely anonymous and data was collected only for a single game session (no longitudinal collection).

DIFFICULTY CHOICE EXPERIMENT

In the chess study [1], players freely chose their opponent with full knowledge of their chess rank. Essentially, this means that players were able to choose the difficulty of their game. In contrast, the study of *Battleship Numberline* [33] involved blindly and randomly assigning players to different levels of difficulty.

Why would knowing the difficulty of a game affect a player's motivation? Knowing the difficulty of a task is likely to affect a player's causal interpretation of their performance. According to Bernard Weiner's attribution theory of motivation [50, 15], this interpretation can have big motivational outcomes. Weiner identified three dimensions of attribution: causality (internal or external), controllability and stability. For instance, if a person attributes their poor performance to low effort (an internal, controllable and unstable cause) they are likely to feel guilty – an emotion that is linked to *increases* in future motivation [50]. In contrast, if they attribute their performance to low ability (which is an internal, uncontrollable and stable cause), they are likely to feel ashamed – an emotion linked to *decreases* in future motivation. Attribution theory provides a theoretical basis for why tasks that are labeled "very easy" might be less motivating than tasks labeled "moderately difficult." If a person is told that a game is "very easy", one is expected to be less proud of their successes, relative to another person who was told that the game was "difficult". Lower pride in one's successes over the course of a game is likely to lower overall task enjoyment. Furthermore, any failure during a "very easy" game might be especially shameful – and shame decreases motivation.

Thus, the inverted-U effect of lowered motivation during tasks labeled "very easy" might be a result of less pride during successes and more shame during failures. According to this theory, the inverted-U in the chess study might have occurred because players generally find it less enjoyable to beat a player they know is weak than to beat a player they know is strong.

Experimental Design: Difficulty Choice Experiment

The primary goal of this experiment is to determine whether giving players a choice of difficulty produces an inverted-U relationship between difficulty and motivation in *Battleship Numberline*. As player choices can be used as a measure of population preferences, we also sought to test whether their pattern of choices resembled an inverted-U shape.

To create five different game levels of difficulty, we used data from a previous experiment [33] and used a regression model to manipulate factors predicted to vary in difficulty from very easy to very hard. We varied several design factors, including Error Tolerance (target size), Time Limit (amount of time players have to make their selection) and Item Sets (items presented).

We randomly assigned players to one of four conditions in a 2x2 between-subjects experiment, as shown in Table 1. Players either received a choice of difficulty, a choice of arbitrary game levels with no information about difficulty, a random level with labeled difficulty, or they were blindly assigned a random level with unlabeled difficulty. Prior to this assignment, players were given a 4-item in-game pretest, which predicts player ability to estimate numbers on a number line, as discussed in [33]. A game session began when players started the game and ended when players exited the game, played more than 80 trials or made no further actions after a time-out. The data presented comes from 10,472 game sessions collected.

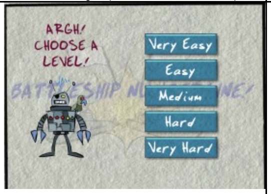
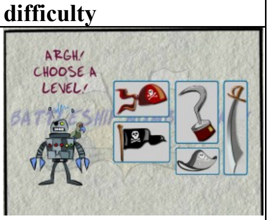
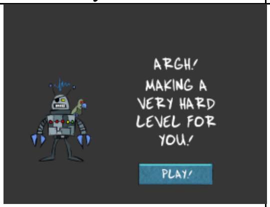
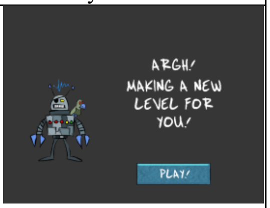
	Information about difficulty (Feedforward)	**No Information about difficulty**
Choice	 ARGH! CHOOSE A LEVEL! Difficulty Choice	 ARGH! CHOOSE A LEVEL! Arbitrary Choice
No Choice	 ARGH! MAKING A VERY HARD LEVEL FOR YOU! Random, No Choice	 ARGH! MAKING A NEW LEVEL FOR YOU! Blind, Random

Table 1: The four conditions in the Difficulty Choice Experiment.

Operational Measures

Difficulty: The average failure rate of a game level or experimental condition.

Engagement: The average of the number of trials played in the game level or condition (total trials divided by total players). A trial is one number line estimate attempt. Engagement, here, is equivalent with intrinsic motivation as there are no extrinsic motivators and players can choose to play another game at any time.

Preference: The tendency for players to select a particular choice (# choices for option X divided by total # choices).

Persistence: Persistence refers to the tendency for players to keep playing the game in the face of failure. Thus, if two students of the same ability were failing at the same rate but one played for longer, we would call the longer playing student more persistent. We measure persistence as a player's actual number of trials minus the predicted number of trials they were expected to play, given their failure rate. A negative persistence score means that players disengaged earlier than the average player.

Results: Difficulty Choice Experiment

To investigate H1.1 ("Providing a choice of difficulty will produce an inverted U-shaped relationship between difficulty and engagement"), we plotted the effect of game difficulty on the duration of player engagement (Figure 2). The inverted U-shaped curve was significant only for players with a choice of difficulty, confirming **Hypothesis 1.1**.

It is notable that, while difficulty choice did produce an inverted-U shape, it did so by depressing motivation on the easy and hard levels, rather than increasing motivation on the moderately difficult levels. This foreshadows the confirmation of **Hypothesis 1.4**, that knowledge of "very easy" and "very hard" levels reduces persistence, and accounts for the inverted-U shape.

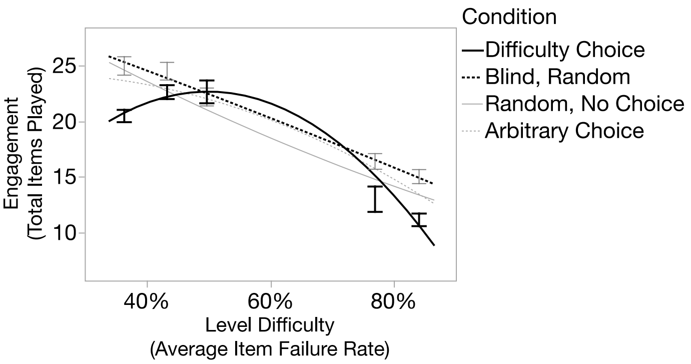


Figure 2: The "Inverted U relation" between difficulty and player motivation is only seen when players can choose their own difficulty condition (solid black line). In comparison to the blind random assignment (dotted black), motivation in the very easy and very hard levels was depressed. Moderate difficulty did not increase motivation, even when self-selected. In all other conditions, the easiest levels were most motivating. The X-axis shows mean failure rate of each of the 5 levels of difficulty, where levels to the right are harder. The Y-axis is a measure of intrinsic motivation (the total number of items players completed). Error bars (standard error) allow for comparison of the means between blind assignment and difficulty choice.

Following [1], our statistical test for an "inverted U-shape" was the significance of the quadratic term (difficulty squared). A squared term in a linear regression tests for curvature in the line of fit; when this term is significant, it indicates significant curvature. We used a response surface regression model of engagement, involving terms for experimental condition, level difficulty, level difficulty squared (the quadratic term) and all interactions. We found that the interaction between condition and level difficulty squared was highly significant (p<0.0001), indicating that the experimental condition caused the curvature of the observed inverted-U. Only the difficulty choice condition had significant curvature.

Figure 3 shows that **Hypothesis 1.2** was not supported ("player choices will resemble an inverted U-shaped curve, with most players choosing moderately difficult levels"). If anything player preference more resembled a U than an inverted U. Players seemed to distinctly prefer the easiest and hardest levels. It is noteworthy, however, that players did not consistently choose to play the easiest games, as might be predicted by the easier is better hypothesis. Figure 3 shows that players preferred the easiest level of play only 32% of the time.

Figure 3 also confirms **Hypothesis 1.3** ("Higher skilled players will self-select themselves into more difficult game levels"). To test this, we first used pretest scores to break players into two equal groups: high and low ability. Players with a high pretest score (i.e., above the median) tended to choose harder levels than players with a low pretest.

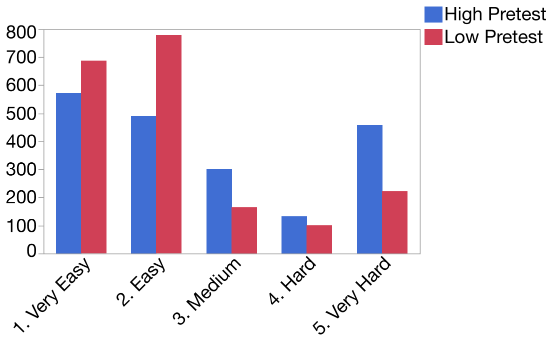


Figure 3: Total count of choices (Y-axis) made for each of the 5 different levels of difficulty, among all players in the Difficulty Choice Condition. The relationship between level difficulty (X-axis) and player preference is not an inverted-U shape – if anything, it is a U shape. Data are evenly split between players with high and low pretest scores.

To test Hypothesis 1.4 ("knowing that game levels are "very easy" will decrease player persistence relative to players who don't know the difficulty of the level they are playing"), we first needed to measure the effect of design variations on persistence, which was defined as how much more or less students were engaged in the face of difficulty. We calculated persistence as the difference between each player's actual engagement (# trials played) and the engagement predicted by a population-level model of the effects of failure on engagement. This linear regression model used data from all players to predict total items played using only their failure rate (total failed items divided by total attempted items).

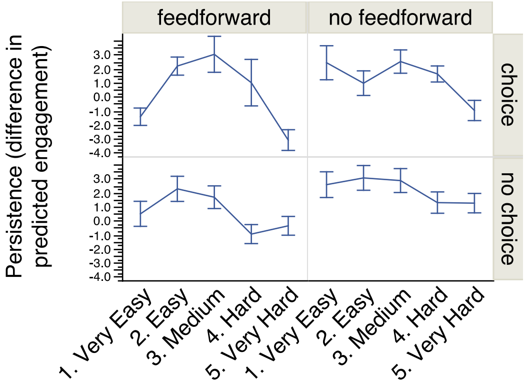


Figure 4: There is reduced player persistence in "Very Easy" and "Very Hard" levels in the "feedforward" conditions (where players know the difficulty). The mere act of labeling levels reduced how long players played, relative to how long they'd be expected to play, given the level's difficulty. Error bars are Standard Error of the Mean.

Figure 4 shows how level difficulty and experimental conditions interacted to affect player persistence. Interestingly, players in the two feedforward conditions (where they were told the level of difficulty) were much less persistent in the "very easy" levels, relative to players who weren't told how easy the levels were, supporting **Hypothesis 1.4**.

Discussion: Difficulty Choice Experiment

When we randomly and blindly assigned difficulty, easier levels were consistently more engaging. But when players were given a choice of difficulty, there was inverted-U relation between difficulty and engagement. Interestingly, this inverted-U was not produced by raising the mid part of the graph, but by lowering the end parts of the graph.

Why might a choice of difficulty make moderately difficult games more motivating? One possibility is that the difficulty and the player's ability are better matched [10]. Another is that increased autonomy improves intrinsic motivation [44,11,13]. A final possibility is that moderately difficult levels *attracted* the most motivated players. It is a limitation of this study that we can't disambiguate these effects. However, this would require mild deception (randomly assigning difficulty irrespective of the choice made by players), which we sought to avoid in the present studies.

Strikingly, players who chose moderately difficult levels did not play longer than players who were blindly assigned to the same moderately difficult levels. For this reason, we suggest that the inverted-U shape emerges due to the predictions of Weiner's Attribution theory of motivation [50]: when players know they are playing a "very easy" game, they get less pride from their successes and more shame from their failures. This leads to less intrinsic motivation (they choose to play fewer items) when doing tasks described as "very easy." This effect appears to depress motivation on games that are known to be very easy – which would otherwise be motivating to players that are unaware of the difficulty level.

One key limitation of these results is that our measure of intrinsic motivation, voluntary engagement (the number of trials players choose to complete), is different from the self-report measure of enjoyment in [1]. While we assume that players choose to play longer because they are enjoying themselves, we can't rule out other reasons. An alternative behavioral measure of enjoyment is preference. Hypothesis 1.2 predicts an inverted U relation between population-level preferences (the choices made by players) and difficulty. As the behavioral choices that people make are indicative of what they enjoy doing, this served as a variation of the inverted-U hypothesis described in [1]. This hypothesis was not supported, as we found that very easy and very hard levels were disproportionately chosen.

Why might this be? One possibility is that primacy and recency effects have influenced the results. This is due to the fact that "Very Easy" was listed at the top and "Very Hard" was listed at the bottom. Murphy *et al.* [39] found that the first item in a list tends to be clicked 19% more than the second item and that the last item in the list tends to be clicked 12% more than the second to last item. Even considering this primacy effect, the preference for very hard levels is still surprisingly strong. Another possibility is that students were just curious about the nature of the very hard difficulty and wanted to explore. Supporting this, other studies have found that interest is a better predictor of student choice of difficulty than player's ability [32].

This first experiment tested whether the non-experimental design used in [1] would produce an "apparent" inverted-U shape in our population. This was confirmed. Note that this experiment was not intended to conclusively show that moderately difficult games cause an increase in player motivation or that they simply attract more motivated players via selection bias. Distinguishing between these two possible reasons for the inverted-U shape would have required an experimental design that involved some level of deceit — we would have had to tell players that they were playing easy levels or hard levels when they weren't. It is notable, however, that the players who were blindly assigned difficulty did not show increased motivation when playing moderately difficult levels.

In summary, when players are randomly assigned game difficulty, easier is better for motivation. When players have a choice of difficulty, moderate levels of difficulty produce the greatest motivation, possibly due to self-selection. Labeling difficulty has the effect of reducing intrinsic motivation for playing very easy games: in the absence of labels, players tend to persist longer in easier games. Finally, the shape of player choices of difficulty is not an inverted U.

Status of the Difficulty Inverted-U Shape

Our evidence does not disprove the inverted-U effect found in the chess study. Perhaps, for instance, our results are idiosyncratic and limited to *Battleship Numberline,* or to educational games, or to games played by novices. These are important limitations. Yet, our evidence is sufficient to question the "truism" that a good game should be neither too hard nor too easy. However, another design factor—novelty—may help explain the face-value importance of a good challenge in a good game.

NOVELTY EXPERIMENT

In the original *Battleship Numberline* study described in the introduction [33], we found just one design factor that increased both difficulty and motivation at the same time: the total number of items presented. A game level that had a small number of easy items was less motivating than a harder game level that had a large number of items. As the small number of items would repeat endlessly, we surmised that the factor of repetition, or its inverse, "novelty", might play an important role in player motivation.

Berlyne [6] conducted a large number of studies on novelty in the 1960s. These studies presented items of varying design (shape, color, size, etc) at various frequencies; the more often items were presented, the more familiar they were—they had less novelty. In general, Berlyne found that subjects increased their rating of pleasantness and interestingness with increased novelty. While these studies dealt with the novelty of individual items, an experience (such as a game) can be said to have more novelty when the design varies more frequently and in more ways. For a recent review of novelty, see [5].

A challenge often combines difficulty with other factors, like novelty. Games with too much difficulty may be frustrating. But is the lack of difficulty, itself, boring? The ratio of positive to negative feedback in most games likely exceeds 10 to 1. Perhaps making games easy is not the cause of the boringness; instead, perhaps easy games fail when they are too repetitive and uninteresting.

A 2016 experiment [26] randomly assigned players to play one of three conditions of "Left 4 Dead 2". In the balanced condition, the number and strength of zombies was normal, in the overloaded condition they were radically increased and in the boredom condition they were decreased to zero. In support of the inverted U effect [1], players reported that the balanced condition was most enjoyable. Additionally, the overloaded condition was significantly more enjoyable than the boring condition. Was the boredom condition less enjoyable because it was so easy (low failure rare) or because there were so few interesting or novel elements in the experience? For instance, when clips of the overloaded condition were shown at CHI16, the audience broke into laughter (there were a LOT of zombies). Perhaps if players had been randomly assigned to simply watch the overloaded condition, despite the lack of difficulty, the condition would still be rated as more enjoyable than the boredom condition.

Interestingly, the construct of "novelty" has also been predicted to create an inverted-U effect on motivation. From The Art of Game Design [46]: "It is impossible to overestimate the importance of novelty as motivation in the realm of game design…Keep in mind, however, that there is such a thing as being too novel." (p.154). This clearly states a testable hypothesis, that a moderate level of novelty will maximize player motivation (Hypothesis 2).

Experimental Design: Novelty Experiment

How to manipulate novelty? Games often introduce novel design changes when players pass into a new "game level". Games with short game levels have a higher frequency of change than games with long levels. When the game itself is controlled, increasing the frequency of change (shorter levels) should increase the rate of experienced novelty.

The following experiment manipulates the frequency of change to deliver different amounts of novelty (frequency of task variation). This is implemented by randomly assigning players to different length game levels. For example, players assigned to game levels that are only 2 items long will have a higher frequency of change than players assigned to game levels that change every 20 items. In this way, we can test the theory that novelty produces an inverted U-shape effect on player engagement, as predicted by game designers and psychologists [46, 49].

In a new version of *Battleship Numberline,* 5,065 players were randomly assigned to six different frequencies of change as six different game level lengths. Players received a new game level after completing a level of 1, 2, 4, 10, 20 or 100 items. The term "game level" refers to a set of trials with a fixed game design (i.e., same time limit, same ship size, etc). Every time a player completed the items in the level, the game declared "Level Up!" and the player received one of 24 randomly selected game configurations, which varied the size of the estimation target, the time limit and the type of target (submarine or ship [33]). After players chose the game, they were presented with a choice of instruction (decimals, fractions or whole numbers). All players then were presented with the 4-item embedded pretest, as discussed in the previous experiment. Any player attempting more than 100 estimation trials was brought back to the menu screen.

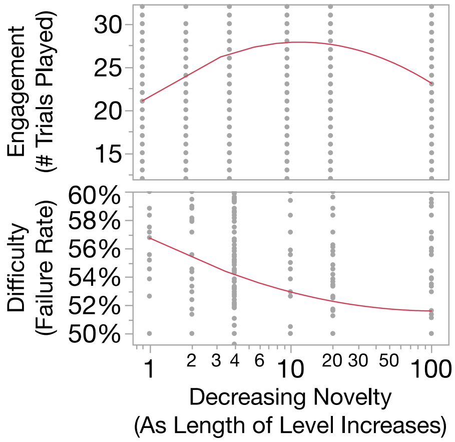


Figure 5: The top graph shows how a change in novelty produces a clear inverted U relation with voluntary engagement (intrinsic motivation measured as number of items played). The X-axis represents decreasing novelty from left to right, as the length of game levels increases (logarithmic scale). The significance of the curvature in this quadratic regression plot is p<0.0001. The lower graph shows how difficulty decreases modestly (p=0.08) as novelty decreases.

Results: Novelty Experiment

Based on the shape seen in Figure 5, the game's novelty (frequency of change) appears to have an inverted-U shape relationship to player engagement (the number of trials they played). Figure 5 shows two quadratic regression plots that model the effects of level length (frequency of change) on the number of trials played and game difficulty. We present the x-axis on a logarithmic scale as that reflects the spacing of the level lengths that we tested (i.e., levels changing every 1 item, 2 items, 4 items, 10 items, 20 items, or 100 items). The quadratic term (level length * level length) is used in order to test the significance of the inverted-U shape [as discussed in 1]. The quadratic of level length was significant (p<0.0001), indicating significant curvature. Therefore, it appears that a moderate level of novelty produces the greatest level of player engagement. Changes in difficulty were not significant (p=0.8).

Discussion: Novelty Experiment

In this experiment, we found that the amount of novelty (task variation) in the game had an inverted U-shaped relationship with player engagement. In other words, our evidence confirms **Hypothesis 2**, the hypothesis that moderate novelty (frequency of change) maximizes player engagement (intrinsic motivation). Berlyne's novelty research [6] found that simple stimuli were more pleasing when novel, but more complex stimuli became more pleasant with more familiarity. Thus, the positive effects of increasing game novelty may depend upon the overall complexity of the game; the novelty effects may be particularly beneficial for simple games. A further limitation of this study design was that it did not independently manipulate novelty and difficulty. As the highest level of novelty is associated with the highest level of difficulty, the negative effects of difficulty may be reducing the observed benefits of novelty. Novelty and difficulty will often be associated with each other, but future work might identify approaches for dissociating novelty and difficulty.

SUSPENSE EXPERIMENT

In the original chess study [1], the researchers identified another factor besides difficulty that generated enjoyment. They found that games were most enjoyable when they were "close games" (the difference between each player's final score was near zero). The authors identified the factor of dramatic suspense (uncertainty [16]) for producing this effect. They noted that sports games where one team beats the other by a wide margin are much less enjoyable for observers than close games. The researchers then followed up on their finding in a separate experiment that manipulated player experience so all participants could experience close games and blow out wins. They again found a considerable effect of the suspense of close games. Given the strength of evidence for suspense, we sought to replicate their findings in our educational gaming context.

Experimental Design: Suspense Experiment

The following experiment investigates the theory that the suspense of a close game will produce an inverted U-shape effect on player motivation, as predicted by [1]. To factor out the role of a player's success/failure, we randomly assigned 6,511 players to receive different standards for winning: players either needed 40%, 60%, 80% or 100% correct in the game level to "win". The present analysis deals with a subset of a larger experiment involving 52,262 play sessions, discussed in [36].

In previous experiments with *Battleship Numberline*, there was no discrete winning or losing state. Players could only succeed at individual tasks (estimating numbers on the number line); there was no explicit "end" of the game or evaluative scorecard.

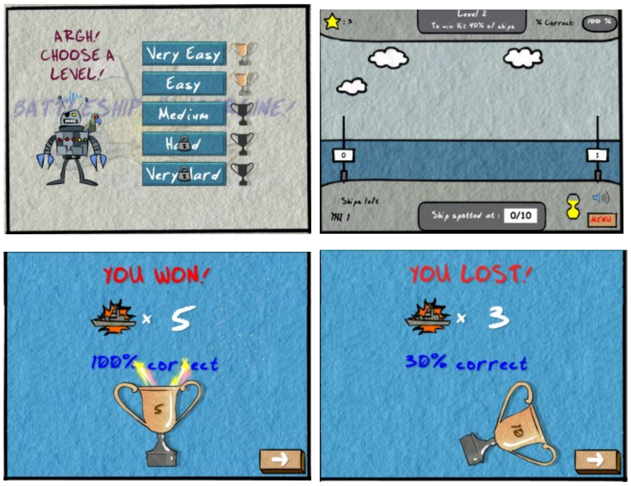


Figure 6: Screens added to support the suspense of winning and losing. The top left screen shows the locking and achievement mechanisms, the top right shows where the goals were displayed to the player and the bottom two screens show the animated screens displaying the winning or losing state.

For this experiment, we added several design elements to a new version of *Battleship Numberline.* First, players were presented with a menu of 5 levels, labeled from Very Easy to Very Hard (as in experiment 1). All levels were locked except for the first. During gameplay, the goal criteria for the level was written at the top of the screen ("to win, hit x% of ships"). After completing either 5 or 10 items, players were shown a scorecard where their score was shown and then they were told if they won or lost. If they won, fireworks were shown. If they lost, the trophy fell over. After pressing continue, players were shown the menu again. If they had won the level, the next level was unlocked and a trophy was shown next to the first level.

Results: Suspense Experiment

To get a measure of the "closeness" of games, we subtracted the game's goal criteria (40%, 60%, 80% or 100%) from the player's success rate. When this closeness was zero or positive, it represented a win; when it was negative, it represented a loss.

Figure 7 shows how this "closeness of game" significantly affects a player's continuing motivation after a win/loss event ("remaining items" refers to the number of items played after the win/loss). Players with the highest positive goal difference score had the highest success rates in the game, however, they were not as motivated to continue playing as players who had barely won. This can be contrasted with all previous experiments, where higher success rates consistently lead to higher engagement (more items played). Figure 7 illustrates the idea that players tend to play for longer when they have a close game in their early levels. Note that a close loss is almost as motivating for continuing play as a blow-out win.

For a statistical test of hypothesis 3, we compared the slope of the line on either side of "0". The "winning" slope was calculated using the distance from 0 (goal difference) as well as the player's actual success rate as factors in a linear regression model. Even after factoring out the player's success rate, the model showed that additional hits beyond what was required to win significantly reduced further play (*p*<0.002); 0.64 fewer items for each 10% increase of score over the win. The same model was then applied to players who lost. In this case, the closer players were to winning, the more they played; 2.3 fewer items less for each 10% decrease in score (*p*<0.0001). As the above slopes are both significant but have opposite valence (-0.64 slope for winners and 2.3 for losers), this is strong statistical evidence for an inverted U-shaped curve—in support of **Hypothesis 3**.

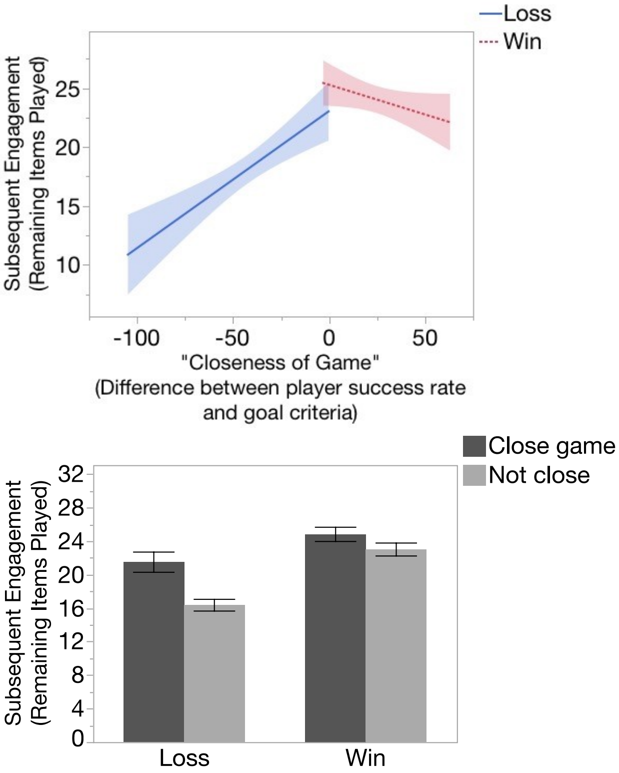


Figure 7: Close games increase player motivation to play, as indicated by the inverted U-shaped curve (top). The X-Axis shows the closeness of the game, or the difference between a player's success rate in the level and the level's goal criteria. Players were randomly assigned to a goal criteria of 40%, 60%, 80% or 100%. Negative closeness indicates that players lost the game. The Y-axis shows the number of items that a player played following the win/loss event ("Remaining Items"). Players with a "blow out win" played significantly fewer additional items than players with a close win. The bottom graph also shows that players experiencing a close loss continued to play almost as many items as players who won.

Discussion: Suspense Experiment

This experimental design attempts to dissociate the effects of winning/losing from the effects of skill. We kept the task difficulty the same over all conditions and only varied, through random assignment, different criteria for winning.

This as successful, in so far as individual player failure rate was not significant in our model of continuing play, yet game "closeness" was. This suggests that the "suspense" of having a close game was the factor creating the inverted U effect seen in Figure 7, rather than the moderate level of difficulty, per se.

Varying the goal criterion is, in a way, similar to varying the difficulty of a task. Thus, are we merely finding that moderately difficult goals create suspense? We suggest that suspense is a different from difficulty and dissociable at either the task or goal level. Suspense can occur at the level of an individual game task or a set of tasks (i.e., a game level). Beyond the suspense of winning, *Battleship Numberline* uses suspense at a task level by providing a slight delay between making an estimate and dropping a bomb at that estimate. This small detail was an explicit part of the original design in order to produce a feeling of suspense by giving the player enough time to wait and discover whether their estimate was successful or not. At a similar task level, Khajah observed that the suspense of a platform-jumping task was dissociable from difficulty, which could be moderated through "covert assistance" [25]. If suspense is a dissociable factor from difficulty, then our findings indicate that there are no motivational benefits from increasing difficulty.

There are several limitations to this study. Ideally, the study design would randomly assign players to win or lose, however this would have involved deception. Instead, we randomly assigned the criteria for winning. However, this means that a "blow-out win" is not possible when the winning criterion is 100%. Furthermore, we reported findings from players during their first attempt at the first level ("Very Easy"). Losing players will necessarily be playing the same level again. This has an unknown effect on their motivation: these players will experience less subsequent difficulty, which can increase motivation, but they will also repeat what they've already done (less novelty), which is expected to decrease their motivation.

GENERAL DISCUSSION

What other factors of challenge, apart from difficulty, might produce an inverted-U shape? We identified three design factors that can generate an inverted-U shaped curve effect: player choice, novelty and suspense. In our three experiments, these three factors appear to be independent from difficulty, suggesting that some of the purported motivational benefits of difficulty may be conflated with other, commonly associated factors.

What is the underlying meaning of the inverted-U shape? Over the years, there have been many descriptions of inverted U-shapes in psychology [reviewed by 49], where a moderate level of some thing (e.g., negative feedback, reward, anxiety, difficulty, complexity, novelty, coffee, etc.) has the effect of maximizing some other outcome (e.g., learning, performance, memory or motivation). To account for these observations, neuroscientist Donald Hebb [19] proposed a general mechanism: he claimed that different situational attributes (such as novelty) could produce an inverted-U shaped effect on performance to the extent that the attributes contributed to *arousal* in the brainstem. Too much or too little arousal will reduce performance; a moderate level of arousal is optimal.

Both novelty and difficulty could potentially affect arousal. However, the present evidence suggests that increased difficulty does not improve motivation. Does an increase in difficulty ever increase intrinsic motivation? To the extent that it does, future work can investigate whether the novelty introduced by the increased difficulty is the primary cause of the increased motivation. For instance, many games get progressively more difficult over time; this is typically viewed as an approach to maintain an optimal level of difficulty as the player's skill increases. An alternative hypothesis is that the changes in difficulty simply help maintain the novelty of the gameplay. According to this hypothesis, the appeal of challenge is more determined by optimal novelty than optimal difficulty.

Limitations

There are a number of important limitations across these experiments. Our findings are specific to a simple, single player, educational, casual game for kids in school; the generalization of these findings to other contexts is, as with all experiments, unknown. The game, *Battleship Numberline*, has the advantage of being simple, which makes it easy to manipulate. However, this simplicity may produce different outcomes (as mentioned in the discussion of the Novelty experiment). Our context, Brainpop.com, has the advantage of being an ecologically valid setting for optimizing player engagement. However, in this context we can only measure the duration of a single session, rather than measuring student progress over multiple sessions. Recent work suggests that difficulty may produce more motivation over a longer time frame [43], which we cannot measure. Our context also does not give us qualitative information about student experience. Another limitation is that our online measures have not been psychometrically validated like other psychological survey instruments, such as the Intrinsic Motivation Inventory [12]. We assume that the measure of "player's total trials attempted" typically reflects a player's free choice to continue the game and not quit. However, we recognize that some data may come from classrooms where children are "forced" to play for instrumental purposes (e.g., a grade). This "noise", however, should not significantly interfere with our results as a whole, due to the random nature of the assignment. We used total trials attempted as a behavioral measure of intrinsic motivation because the use of total time (in seconds) has many more extreme outliers and the process of removing these outliers is prone to introduce bias. Better instrumentation could resolve this in future experiments.

We recognize that large-scale "real-world" environments will produce much noisier data than laboratory experiments.

One effect of this scale and noise is that data can be easily manipulated to show different effects. We encourage healthy skepticism. In this paper, we aimed to present simple data stories, tried to avoid complex statistical techniques and took the approach that our findings should be robust in the face of different approaches to analyses (e.g., with different data filters). One final measurement limitation of this work is that we did not analyze learning curves across conditions. Previous work has found that faster learning occurs with more difficult conditions [33], though this effect is questionable [43]. Thus, the findings here can only inform theories regarding the relationship between difficulty and motivation, not the subsequent effects on student learning, which remains for future work.

Additional Future Work

All data sets are available for secondary analysis at the PSLC Datashop [28]. Our findings are primarily directed at the design of educational games, but future work can extend to entertainment games. Tom Malone's seminal work with game design factor analysis [35] is a model for future online game experiments; the entire taxonomy of intrinsic motivations for learning [38] represents testable hypotheses. Future work can deconstruct game challenges into their underlying functional factors, such as novelty, difficulty and others. As some games (e.g., Dark Souls and Flappy Bird) engage users primarily through excessive failure, it may be productive to explain these unusual failure-oriented games with the hypothesis that excessive difficulty is used a mechanism for providing novel player experiences.

Design Implications

While the notion of challenge has many positive connotations, "difficulty" directly refers to the potential for task failure [27,33]. These two terms are useful to distinguish. Our evidence implies that early game experiences should minimize difficulty and provide a moderate degree of novelty. This does not mean designers should avoid challenge. Instead, designers should ensure that novice players receive significant amounts of positive feedback during challenges. When playing something new, players generally like to feel successful and competent. Task repetition causes fatigue and should be accompanied by a regular drip of novelty. After a time, difficulty itself can provide this source of novelty. Both low and high performance players benefit from a feeling of suspense during a close game. In general, challenge appears to be fun because it is interesting. Keeping games easy and interesting may be more important than "balancing difficulty." For this reason, we suggest the maxim "not too hard, not too easy" might be restated as "not too hard, not too boring."

CONCLUSION

Our three experiments investigated how different challenge factors (difficulty, choice, novelty and suspense) affected intrinsic motivation. Our goal was to identify conditions where these factors of challenge might produce an inverted U-shaped effect on intrinsic motivation, as predicted by Flow Theory [10]. In summary, we found that providing players with a choice of difficulty produced an inverted-U shaped relationship between difficulty and motivation, primarily by depressing motivation on very easy levels. We also found that suspense (close games) and balanced novelty increased player motivation. However, none of our experiments showed that difficulty, by itself, actually *caused* improved motivation. Within our experimental context, our evidence indicates that increasing difficulty consistently reduces motivation, when other motivational factors are controlled. Recognizing the richness of the concept of challenge and its role in game design, these other components of challenge may provide the motivational benefits long attributed to difficulty.

We note that motivational theory does not always generalize from the laboratory to the real world. How can we identify and develop theories with strong external validity and the capacity to broadly generalize? Massive online experiments may be useful for searching the space of circumstances under which a theory will or will not apply [29]. Whether or not a scientific theory will generalize to different types of games or different types of players will always be an empirical question.

There seems to be a vast scientific and practical benefit that can be realized from running large-scale experiments inside of real products. As practitioners become increasingly comfortable running A/B tests to evaluate different designs [30], we encourage designers of popular games (educational or otherwise) to conduct more experiments designed to produce generalizable findings. This kind of basic research can benefit designers [35] by improving design theory.

The sheer scale of online gaming, which involves millions of diverse participants, could be a source for experiments that could significantly inform our scientific understanding of human motivation and design. It is promising to consider that game design patterns [7], principles [46,14] and exemplars embody hundreds of implicit and explicit hypotheses about human motivation. The sciences of motivation and design might be rapidly advanced if these design patterns were linked to psychological theory and systematically tested online.

ACKNOWLEDGMENTS

Special thanks to BrainPop and Allison Levy for their support and encouragement! The research was supported by Carnegie Mellon's Program in Interdisciplinary Education Research (PIER) funded by grant number R305B090023 from the US Department of Education and the Institute of Education Sciences through GrantR305C100024 to WestEd and by DARPA ONR N00014-12-C-0284. Additional thanks to the UCSD Design Lab community, Mike Mozer, Mohammad Khajah, Jesse Schell, Gary and Judy Olson, the HCIC community, Scott Klemmer, Don Norman, Erik Harpstead, Bob Kraut, Playpower Labs and Kishan Patel.

REFERENCES

1. Abuhamdeh, S. and Csikszentmihalyi, M. (2012) The Importance of Challenge for the Enjoyment of Intrinsically Motivated, Goal-Directed Activities. *Personality and Social Psychology Bulletin 38*, 3, 317–330.
2. Andersen,E.,Liu,Y.,Snider,R.,Szeto,R.,and Popovic, Z. (2011) Placing a value on aesthetics in online casual games. *ACM CHI.*
3. Aponte, M., Levieux, G., & Natkin, S. (2011). Measuring the level of difficulty in single player video games. *Entertainment Computing*, *2*(4), 205–213.
4. Atkinson, J. (1958) 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, 288–305.
5. Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or surprise? *Frontiers of Psychology*, 4
6. Berlyne, D. E. (1970). Novelty, complexity, and hedonic value. *Perception & Psychophysics*, *8*(5), 279-286.
7. Bjork, S., & Holopainen, J. (2004). *Patterns in game design*.
8. Chase, C.C. (2012). The interplay of chance and skill: Exploiting a common game mechanic to enhance learning and persistence. *Proceedings of the 2012 International Conference of the Learning Sciences.*
9. Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715.
10. Csikszentmihalyi, M., & Csikszentmihalyi, I. S. (Eds.). (1992). *Optimal experience: Psychological studies of flow in consciousness*. Cambridge University Press.
11. 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.
12. Deci, E. L., & Ryan, R. M. (2003). Intrinsic motivation inventory. *Self-Determination Theory*, *267*.
13. Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, *125*(6), 627.
14. Dickey, M.D. (2005) Engaging by design: How engagement strategies in popular computer and video games can inform instructional design. *Educational Technology Research and Development 53*, 2, 67–83.
15. Eccles, J.S. and Wigfield, A. Motivational beliefs, values, and goals. *Annual review of psychology 53*, (2002), 109–32.
16. Ely, J., Frankel, A., & Kamenica, E. (2015). Suspense and surprise. *Journal of Political Economy*, *123*(1), 215-260.
17. Fishbach, A. and Finkelstein, S. How feedback influences persistence, disengagement and change in goal pursuit. In A. Elliot and H. Aarts, eds., *Goal-directed behavior*. Psychology Press, 2012, 1–53.
18. Harter, S. (1978a) Pleasure derived from challenge and the effects of receiving grades on children’s difficulty level choices. *Child Development 49*, 3 788–799.
19. Hebb, D. O. (1955) Drives and the C.N.S. (Conceptual Nervous System) *Psychological Review*, *62*, 243-254
20. Van der Heijden, H. (2004). User acceptance of hedonic information systems. *MIS Quarterly*, *28*(4), 695–704.
21. Hockley, R. (2013) *The Psychology of Fatigue: Work, Effort and Control.* Cambridge University Press.
22. Hong, Y., Chiu, C., Dweck, C.S., Lin, D.M.-S., and Wan, W. Implicit theories, attributions, and coping: A meaning system approach. *Journal of Personality and Social Psychology 77*, 3 (1999), 588–599.
23. Hunicke, R., & Chapman, V. (2004). AI for dynamic difficulty adjustment in games. *Challenges in Game Artificial Intelligence AAAI*, 91–96.
24. Kashdan, T. B., & Silvia, P. J. (2009). Curiosity and interest: The benefits of thriving on novelty and challenge. *Oxford handbook of positive psychology*, *2*, 367-374.
25. Khajah, M. M., Roads, B. D., Lindsey, R. V, & Mozer, M. C. (2016). Designing Engaging Games Using Bayesian Optimization. *ACM CHI.*
26. Klarkowski, M., Johnson, D., Wyeth, P., McEwan, M., Phillips, C., & Smith, S. (2016). Operationalising and Evaluating Sub-Optimal and Optimal Play Experiences through Challenge-Skill Manipulation. *ACM CHI.*
27. Kline, P. (2014). *The new psychometrics: Science, psychology and measurement*. Routledge. p.79
28. Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. *Handbook of educational data mining*, *43*.

29. Koedinger, K. R., Booth, J. L., Klahr, D. (2013) Instructional Complexity and the Science to Constrain It *Science*. 22 November 2013: Vol. 342 no. 6161 pp. 935- 937

30. Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R.M. (2008) Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery 18*, 1 140–181.

31. Kurzban, R., Duckworth, A., Kable, J.W., and Myers, J. An opportunity cost model of subjective effort and task performance. *The Behavioral and brain sciences 36*, 6 (2013), 661–79.

32. Inoue, N. (2007). Why face a challenge?: The reason behind intrinsically motivated students' spontaneous choice of challenging tasks. *Learning and Individual Differences*, *17*(3), 251–259.

33. Lomas, D., Patel, K., Forlizzi, J. L., & Koedinger, K. R. (2013) Optimizing challenge in an educational game using large-scale design experiments. *ACM CHI*

34. Lomas, D., Forlizzi, J., Poonwala, N., Patel, N., Shodhan, S., Patel, K., Koedinger, K., and Brunskill, E. (2016). Interface Design Optimization as a Multi-Armed Bandit Problem. *ACM CHI.*

35. Lomas, D. (2015) Accelerating Theory Development with Large Online Experiments: Towards an Interaction Design Science. *HCIC*

36. Lomas, J. D. (2014). *Optimizing motivation and learning with large-scale game design experiments* (Doctoral dissertation, Carnegie Mellon University).

37. Malone, T. (1981) Toward a theory of intrinsically motivating instruction. *Cognitive Science 5*, 4, 333-369.

38. Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. *Aptitude, learning, and instruction*, *3*(1987), 223-253.

39. Murphy, J., Hofacker, C. and Mizerski, R. (2006) *Primacy and recency effects on clicking behavior.* Journal of Computer Mediated Communication, 11 (2).

40. Norman, D. A. (2013). *The design of everyday things: Revised and expanded edition*. Basic books.

41. O'Rourke, E. and Haimovitz, K. (2014) Brain points: a growth mindset incentive structure boosts persistence in an educational game. *ACM CHI.*

42. Papoušek, J., & Pelánek, R. (2015) Impact of Adaptive Educational System Behaviour on Student Motivation. AIED

43. Pelánek, R., Rihák, J., & Papoušek, J. (2016). Impact of data collection on interpretation and evaluation of student models. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 40-47). ACM.

44. Shalley, C. E., & Oldham, G. R. (1985). Effects of goal difficulty and expected external evaluation on intrinsic motivation: A laboratory study. *Academy of Management Journal*, *28*(3), 628-640.

45. Shapira, Z. (1989) Task choice and assigned goals as determinants of task motivation and performance. *Org. Behavior & Human Dec. Proc.* 44, 2, 141–165.

46. Schell, J. (2008) *The Art of Game Design.* Morgan Kaufmann

47. Schmierbach, M., Chung, M., Wu, M., & Kim, K. (2014). No One Likes to Lose: The Effect of Game Difficulty on Competency, Flow, and Enjoyment. *Journal of Media Psychology*, *26*(3), 105–110.

48. Sweetser, P. and Wyeth, P. (2005) GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE) 3*, 3, 1–24.

49. Teigen, K. (1994) Yerkes-Dodson: A Law for all Seasons *Theory & Psychology* 4:525-547,

50. Weiner, B. An attributional theory of achievement motivation and emotion. *Psychological review 92*, 4 (1985), 548–73.

51. Yannakakis, G.N., Denmark, S., and Hallam, J. (2007) Towards optimizing entertainment in computer games, 933-971

52. Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and psychology*, *18*(5), 459-482.