Meet your Match Rating: Providing Skill Information and Choice in Player-Versus-Level Matchmaking

Anurag Sarkar
Northeastern University
sarkar.an@husky.neu.edu

Seth Cooper
Northeastern University
scooper@ccs.neu.edu

ABSTRACT

Previous work has demonstrated the effectiveness of rating system-based matchmaking for level ordering within the particular constraints of human computation games. However, players were not informed about the rating system, nor allowed to choose the difficulty of upcoming levels. Informing players of the ratings used in the system and offering them choice of upcoming level difficulty may enhance feelings of competence and control respectively, thereby further improving player engagement. Thus, we attempted to improve player experience by both exposing players to the underlying rating system, as well as offering them choice of level difficulty. We found that players cognizant of ratings both attempted and completed more levels than those who were not. Though additionally offering choice did not significantly affect behavior, we found that player choice was influenced by the outcome of the preceding level. Moreover, we did not observe any significant impact on self-reported measures of subjective experience.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**;

KEYWORDS

human computation games; ratings feedback; player choice; player engagement

ACM Reference Format:
Anurag Sarkar and Seth Cooper. 2018. Meet your Match Rating: Providing Skill Information and Choice in Player-Versus-Level Matchmaking. In *Foundations of Digital Games 2018 (FDG18), August 7–10, 2018, Malmö, Sweden.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3235765.3235795

1 INTRODUCTION

Human computation games (HCGs) attempt to solve computationally intractable problems by modeling them as games that aim to leverage the skills of large numbers of human players. HCGs have found success in tasks such as protein folding [8], image labeling [38] and software verification [6, 11] among others.

Such success notwithstanding, HCGs present unique challenges to achieve difficulty balancing, due to the constraints imposed by having to solve pre-existing problems that cannot be freely modified and whose difficulties are often unknown in advance. Thus, game levels which are based on these unsolved, real-world problems cannot be manipulated or reasoned about *a priori* as a means to implement difficulty balancing. To this end, rating systems such as Elo [14], Glicko/Glicko-2 [17, 18] and Microsoft's TrueSkill [19] were suggested [7] as a possible means to overcome these difficulty balancing barriers within HCGs. Later work [32] demonstrated this approach to be effective: by giving both players and levels Glicko-2 ratings, performing matchmaking between players and levels was used to manipulate the ordering of levels served to players rather than manipulating the levels themselves.

However, while past work was able to demonstrate the effectiveness of using rating systems in balancing difficulty and improving engagement within HCGs, it did so while keeping players out of the loop—players were not aware of their in-game performance and skill level while playing through only those levels that the system served them. Though the matchmaking algorithm was shown to be effective in serving players levels with challenge appropriate to their skill, additionally offering players information about the ratings system as well as giving them the ability to select their desired degree of difficulty may further improve engagement by tapping into additional motivational factors.

This work builds upon previous work [32] by examining the effects of exposing the players to the underlying Glicko-2 rating system used to perform level ordering and matchmaking in the HCG *Paradox* and also studying the effects of offering players the choice in difficulty of each level. Although we found no observable impact on self-reported measures of subjective experience, we found that players who were aware of their ratings and the rating system spent significantly more time playing and also attempted and completed significantly more levels than those who were not. Moreover, both informed and uninformed players completed levels of similar difficulties. While additionally offering choice of difficulty seemed to not significantly impact player experience or engagement, we noticed the choice made by players to be influenced by their outcome on the preceding level.

This work contributes an empirical study demonstrating that 1) informing players about the underlying system used for matchmaking can improve engagement in human computation games and 2) player's selection of the difficulty of an upcoming level is influenced by the outcome of the previous level.

2 BACKGROUND

This work draws on a wide range of background literature that relates to informing players of their ratings and giving them choice in difficulty as means to improve engagement.

FDG'18, August 7-10, 2018, Malmö, Sweden

ACM ISBN 978-1-4503-6571-0/18/08…$15.00
https://doi.org/10.1145/3235765.3235795

2.1 Engagement and Flow

Engagement is a concept in the psychology of motivation and play that attempts to capture how involved a person feels in a task that they are performing. This is related to Csikszentmihalyi's theory of flow [10] which talks about achieving the "flow state"—an optimal state of mind where an individual is deeply engrossed in an activity and is motivated to do it well. Understandably, most games aim to engage players and push them towards this flow state in order to deliver optimal player experiences. Traditional methods of optimizing engagement have involved difficulty balancing, i.e. having in-game challenges be tailored to the specific player's skill level [1, 5, 12, 15, 16, 24, 26], though recent work has suggested that other factors, such as novelty, might play a bigger role in enhancing player engagement [27]. While previous work [32] has shown that player engagement within HCGs can be increased by implementing such difficulty balancing with the help of player-versus-level matchmaking using the Glicko-2 rating system [18], further improvements in engagement may be achieved by offering players additional motivational factors such as competence and choice.

2.2 Self-Determination Theory

Informing players about their ratings and giving them choice may additionally help satisfy specific player needs as identified by self-determination theory (SDT) [30, 31]. SDT is a theory of motivation that argues that the quality of motivation experienced by an individual is governed by three innate psychological needs, namely, *autonomy*, *competence* and *relatedness*.

Ryan and Deci state that while *intrinsic* motivation ("the doing of an activity for its inherent satisfactions rather than for some separable consequence" [30]) satisfies autonomy and competence, *extrinsic* motivation ("'[the doing of an activity] in order to attain some separable outcome" [30]), due to not being inherently interesting, must satisfy a sense of belonging and connection to people, groups or cultures, for the purpose of achieving a certain goal. This sense is referred to as *relatedness*. Thus, SDT claims that conditions that satisfy the needs for autonomy, competence and relatedness, lead to individuals experiencing enhanced feelings of engagement, motivation and performance.

HCGs, due to their inherent nature of harnessing the collective ability of large groups of individuals to solve important problems, may already satisfy player needs for *relatedness*. By informing players of their skill level and offering them choice of level difficulty, we may also satisfy the needs for *competence* and *autonomy* respectively, and thereby achieve both measurably increased engagement as well as improved player experience.

Thus, we wanted to determine if our experimental design helped increase engagement by leveraging SDT. Several questionnaires exist to assess the different constructs that are present within the theory. For our experiment, we decided to use the *Intrinsic Motivation Inventory (IMI)* [31] which is designed for examining the subjective experience of participants who work on an interesting task within the bounds of different experimental conditions, which describes our study quite well. The IMI can be used to gauge the participants' intrinsic motivation via the subscales of interest/enjoyment, perceived competence, perceived choice, effort/importance, value/usefulness and felt pressure and tension, while they undertake a given task. For our experiment, we used the first four of these subscales, which we describe in more detail in section 3.4.

2.3 Dynamic Difficulty Adjustment

Dynamic difficulty adjustment (DDA) refers to a general category of techniques for dynamically altering the in-game degree of difficulty in response to player performance. Several such techniques have been employed in past work, ranging from tweaking in-game parameters [20] and modifying level design (such as in *Left 4 Dead* [9]), to procedurally generating level segments using machine learned models of difficulty [21] and ordering levels generated by users using the *TrueSkill* rating system on player attempt outcomes, as in the platformer *JumpCraft* [36].

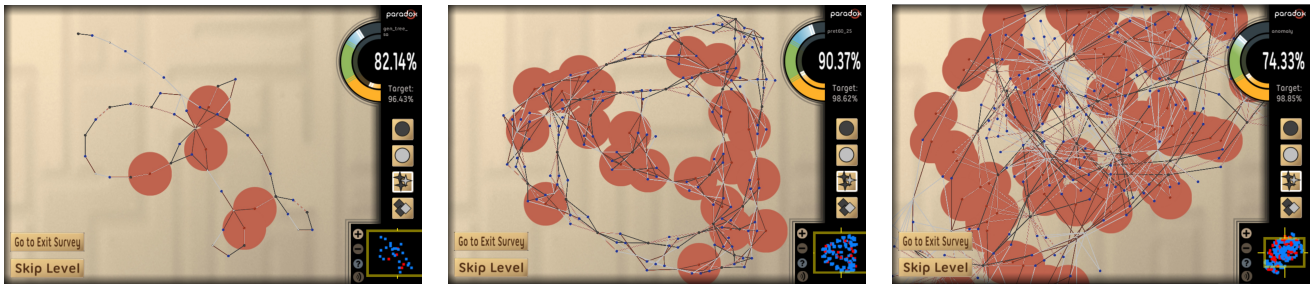
Baldwin et al. [2, 3] studied the effects of informing players about the existence of DDA techniques within the game, specifically focusing on multiplayer DDA. They found that giving players this additional information reduced the effectiveness of the DDA system since skilled players were better able to exploit knowledge about the system to gain an advantage over those with less skill. Since HCGs are often non-competitive, it is likely that informing players about the workings of any underlying difficulty adjustment mechanism would not necessarily reduce its effectiveness as in the multiplayer setting that Baldwin et al. examined. Hence, DDA techniques might still benefit from player feedback which may help in guiding the underlying difficulty adjustment algorithms at work, particularly in a non-competitive HCG setting like ours.

2.4 Feedback

Exposing the players to their ratings and the underlying rating system can be considered a form of feedback. It has long been known that appropriate feedback can support enjoyment and learning, often examined in educational contexts [4, 33, 37]. Recent work has also shown that giving feedback to workers in a crowdsourcing context can be beneficial [13].

Siu et al. [34] studied the effects of reward feedback within the specific context of HCGs. Their findings showed that offering players choice of reward between multiple reward systems had positive effects on both task completion and player experience. Giving players such choice of reward seems to simultaneously tap into the needs for both competence and autonomy, leading to higher player engagement. This suggests that such benefits may also be observed by satisfying these needs with the help of choice of difficulty (rather than choice of reward) and skill-based feedback (rather than reward feedback) by exposing players to the rating system, as in our experiment.

Mekler et al. [29] studied the effects of feedback in the form of points, levels and leaderboards within an image annotation task. Similar to [34], the study showed that feedback helped increase the rate of task completion (via increased image tag quantity) but failed to positively impact tag quality, intrinsic motivation or need satisfaction. These findings indicate that feedback of this form acts as extrinsic (rather than intrinsic) motivation and is suitable for



Rating: 961 (easy) Rating: 1476 (medium) Rating: 1739 (hard)

Figure 1: Example levels from our data set covering a range of ratings.

producing increased task completion, since though intrinsic motivation was not augmented, it was not negatively impacted either. While our ratings-based skill feedback is similar to the feedback studied by Mekler et al., additionally offering choice within our HCG tasks could positively impact intrinsic motivation while still preserving the task completion benefits that skill feedback was shown to offer.

All of this feedback-based literature informed the formulation of our first hypothesis which claimed that giving ratings-based feedback to players would improve their engagement over not giving them such feedback.

2.5 Difficulty Choice

Several classic and modern games offer players the ability to choose their preferred level of in-game difficulty. The traditional and most prevalent form of doing so is to present users with the static difficulty options of "Easy", "Normal" and "Hard" at the start of the game, but game difficulty choices can take many other forms and modalities such as biofeedback, for example [25].

Smeddinck et al. [35] explored how offering different modes of difficulty choices affect player experience. Their findings suggest that players exhibit a preference for manually choosing in-game difficulty which informed our decision to offer players the ability to manually choose whether the next level in the game should be of a low, high or recommended difficulty. Moreover, Smeddinck et al. found that embedding difficulty choices within games do not significantly impact game experience apart from perceived autonomy. Thus, we examined if this increased autonomy afforded by manual choice, combined with competence provided by revealing the rating system and relatedness offered by HCGs, helped improve engagement. This led to our second hypothesis, where we claimed that offering such difficulty choice in addition to revealing the ratings to players, would further increase player engagement over only revealing the ratings and not giving choice.

3 METHOD AND SYSTEM

3.1 Hypotheses

Based on the literature, we formulated the following hypotheses:

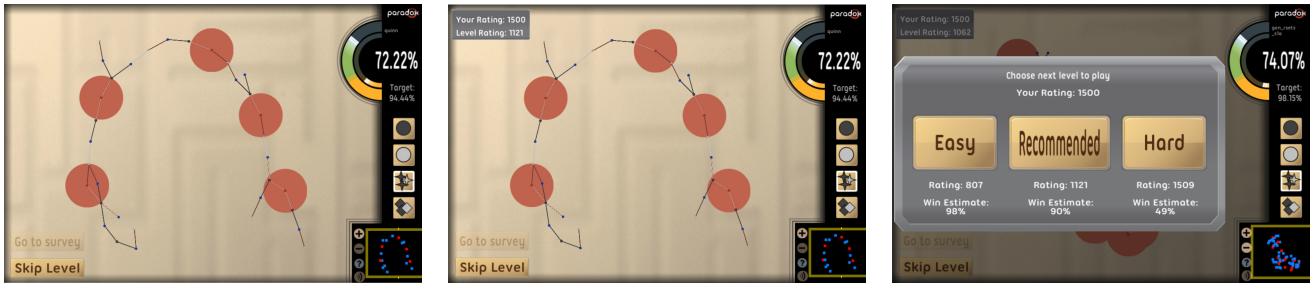
- *H1*: Informing players of their rating (as well as explaining the rating system and encouraging them to get a high rating) will lead to higher behavioral engagement and higher self-reported measures of experience than not informing the players about their ratings or the underlying rating system.
- *H2*: Additionally giving players choice when informing them of their ratings will lead to even higher measures of both behavioral engagement and self-reported experience than those observed when informing them of the rating system but not offering them choice.

3.2 Game Description

For our study, we used *Paradox* [11], a 2D puzzle HCG originally designed for crowdsourced formal verification where each level within the game corresponds to a maximum satisfiability (MAX-SAT) problem. The levels are graph-like structures with the vertices corresponding to variables and clauses in the underlying MAX-SAT problem. Players can utilize both manual as well as automated tools (represented as "brushes") to color these vertices, in turn assigning values to the variables in the underlying problem, in order to satisfy as many clauses as possible. The player's score for each level corresponds to the percentage of clauses that the player is able to satisfy. A player 'completes' or finishes a level by reaching a pre-determined target score. Each level thus corresponds to one specific human computation task whose difficulty is estimated as the level's Glicko-2 rating. *Paradox* has found application in past studies [32, 39] of engagement within HCGs and thus proved to be a suitable testbed for this work. Example levels of different difficulties and their corresponding ratings are shown in Figure 1.

The version of the game in our experiment consisted of 9 hand-authored tutorial levels served in a fixed order that had to be completed, followed by 55 optional challenge levels whose order was customized by a ratings-based matchmaking system. Of the 55 challenge levels, 17 were generated by us using randomized algorithms for SAT problem generation, while the remaining 38 were drawn from the set of SATLIB Benchmark Problems[1]. When served a level, a player had the option to skip the level (ignored for the purposes of rating); once they made a move, the player could either complete the level by reaching a target score, or forfeit the level (interpreted as a win or loss, respectively, for ratings). Ratings for the levels were generated using match data from past HITs involving *Paradox* [32, 39]. For this experiment, the level ratings obtained ranged from 807 to 2276 on the Glicko-2 scale.

[1] http://www.cs.ubc.ca/~hoos/SATLIB/benchm.html



BLIND RATINGS CHOICE (after a level)

Figure 2: Screenshots for each condition.

3.3 Participant Recruitment

For recruiting players, we framed our experiment as a Human Intelligence Task (HIT) posted on the crowdsourcing platform Amazon Mechanical Turk (MTurk). Though MTurk participants are paid workers, studies have shown that these workers are motivated by enjoyment in addition to payment [22, 28]. Past studies have also successfully used paid recruitment through MTurk to examine voluntary engagement in games [23, 32].

The details of our HIT were as follows:

> *Title: Human Computation Puzzle Game*
> *Description: Play a puzzle game derived from a real-world problem. You would need Adobe Flash Player 10.0 or greater to proceed.*
> *Keywords: survey, game, play, puzzle*

The HIT paid $1.75 and indicated that the expected time to finish the task was 30 minutes. Note that this was the estimated time to play through the *entire* game (i.e. the 9 tutorial levels along with the 55 challenge levels), which was optional. Workers only needed to complete the tutorial levels and skip through 5 challenge levels, and then complete the survey. These actions took approximately 7, 1, and 1 minute respectively. Adding a minute to read instructions, the players' required actions took roughly 10 minutes, making the average pay rate about $10.50 per hour. In our experiment, the workers spent a median time of 6.9 minutes on the tutorial levels, and a median time of 11.1 minutes on the challenge levels.

Players were given the following instructions:

> *There are three stages to the HIT:*
> 1. *Play and complete all the tutorial levels.*
> 2. *Try to complete as many challenge levels as you can!*
> 3. *Go to the survey and complete it.*
>
> *You MUST complete all the tutorial levels. The survey will not be accessible during the tutorial and will become available once you fail to complete (i.e. skip/forfeit) at least FIVE challenge levels*
> *It is NOT necessary to complete all challenge levels and you will be given the completion code as long as you complete the survey.*

For our experiment, players were required to complete all nine tutorial levels in order to become familiar with the game. These levels were used only to help the player understand the mechanics of the game and data from these levels were not taken into account for our analyses.

Upon completing the tutorial phase, players proceeded to the challenge phase, which consisted of a total of 55 different levels served dynamically, as described above. After skipping or forfeiting a total of 5 levels, the player was additionally offered the option to finish playing at any time and go directly to the post-game survey. Once a level was seen by a player, that level was removed from the level pool, thereby allowing each player to see each level only once. Although no player played through all 55 levels in the game, if this were to have happened, the player would have been informed that there were no more levels left and taken directly to the post-game survey.

After finishing the game, the players completed the *Intrinsic Motivation Inventory* questionnaire. The version we presented consisted of 25 questions that measured the subscales of *Interest/Enjoyment*, *Perceived Competence*, *Perceived Choice* and *Effort/Importance*. The HIT was completed upon submitting this survey.

3.4 Ratings Feedback and Choice Experiment

For each participant, we measured behavioral engagement by tracking the following variables:

- *Challenge Time*: The total time in seconds spent by the player in the challenge levels.
- *Levels Attempted*: The number of levels *attempted* by a player, where they made at least one move.
- *Levels Completed*: The number of levels *completed* by a player, where they reached the target score.
- *Player Rating*: The player's rating upon completing the HIT.
- *Highest Level Rating*: The highest rating of any completed level (set to 0 if a player didn't complete any levels).

In addition to the above metrics for engagement, we also conducted a post-game Intrinsic Motivation Inventory (IMI) survey[2] in order to test our hypotheses through the lens of self-determination theory, as described in the background section. We used the following subscales of the IMI in order to examine the players' self-reported measures of subjective experience.

- *Interest/Enjoyment (scale of 7 to 49)*
- *Perceived Competence (scale of 6 to 42)*
- *Perceived Choice (scale of 7 to 49)*
- *Effort/Importance (scale of 5 to 35)*

All questions for each subscale were used. Each question was scored from 1 to 7 and subscales had 5 to 7 questions, resulting

[2] http://selfdeterminationtheory.org/intrinsic-motivation-inventory/.

in the above ranges for each subscale. The *Interest/Enjoyment* subscale in particular is considered to be the self-reported measure of intrinsic motivation. Moreover, the subscales of *Perceived Competence* and *Perceived Choice* are utilized as additional positive predictors of behavioral and self-reported intrinsic motivation. The *Perceived Choice* subscale was particularly relevant in this work on account of our choice-based experimental condition. Finally, the *Effort/Importance* subscale gives a self-reported measure of how much effort the player put into the game.

Our experiment consisted of three different conditions. These were: serving levels without informing players of either their own or the level's ratings (BLIND), serving levels while informing the player about their own rating, the level's ratings as well as briefly explaining how the rating system works (RATINGS), and serving levels as in the second condition but additionally offering the players the choice of the next level being of easy, recommended or hard difficulty based on their current ratings (CHOICE).

For the first two conditions, the order of levels served was determined by using a matchmaking algorithm which takes as input the player's current Glicko-2 rating and is thus based on the player's in-game performance. Similar to previous work [32, 39] that used matchmaking in *Paradox*, the player's desired win probability was computed based on their current Glicko-2 rating. The desired win probability is the win probability versus levels that we wanted players to have based on current skill estimates and is discussed further in [32]. Additionally, we also computed the player's expected win probability against each level using Glicko-2's E function. The level that was then served was the one against which the player's expected win probability was closest to the player's desired win probability. In this way, the matchmaking algorithm differed from the one used in [32] where the level served was randomly selected from a group of levels whose expected win probabilities were within a certain window of the player's desired win probability. Additionally, when determining the level to be served, we removed from consideration the easiest (i.e. lowest rated) and hardest (i.e. highest rated) remaining levels because we wanted to ensure that there would always be at least one level easier than and one level harder than the recommended level for the CHOICE condition.

In the CHOICE condition, we applied the matchmaking algorithm as in the first two conditions to determine the level for the recommended option, but used as input the player's current rating $r - 400$ to determine the level for the easy option, and the player's current rating $r + 400$ to determine the level for the hard option. In other words, if a player only ever selected recommended levels in the CHOICE condition, then she would receive the same order of levels as if she were in the RATINGS or BLIND condition. The $+/-400$ value was based on the formula for expected win probability, and roughly corresponds to an order of magnitude difference in the win probability of the player or level.

In all conditions, after each match, the player's rating was updated using the Glicko-2 system and the whole process was repeated for each match the player was involved in until the player exited the game by going to the survey.

Additionally, prior to the challenge levels section, in the BLIND condition, players were given the following instructions:

Variable	BLIND	RATINGS	CHOICE
Challenge Time†	515^{a}	791^{b}	897^{b}
Levels Attempted†	7^{a}	10^{b}	12^{b}
Levels Completed†	5^{a}	7^{b}	8^{b}
Player Rating	1500	1500	1525
Highest Level Rating	1222	1293	1413
Interest/Enjoyment	63%	65%	63%
Perceived Competence	57%	52%	57%
Perceived Choice	78%	80%	82%
Effort/Importance	83%	86%	83%

Table 1: Summary table of variable analysis. Median values are provided for all variables, but represented as percentages of maximum possible obtainable value for survey variables. Variables with daggers† had significant differences in omnibus tests. Values with differing superscripts^{a,b} had significant differences in post-hoc tests. More detail on statistical tests is given in Table 2.

> *Use the skills you've learned to play the upcoming challenge levels. You now have the option to skip levels, or go to the survey if you wish.*

Alternately, in the RATINGS and CHOICE conditions, where the players were informed of their ratings, they were given the following instructions:

> *Use the skills you've learned to play the upcoming challenge levels. You now have the option to skip levels, or go to the survey if you wish. Gameplay is regulated by a rating system. Each challenge level is assigned a rating indicating its difficulty. The higher the rating, the harder is the level. You are assigned a default starting rating of 1500. When you complete a level, your rating goes up, and when you forfeit a level, your rating goes down. Your rating is unaffected when you skip a level. Try to get as high a rating as you can!*

Thus, both informed and uninformed players were encouraged to complete as many levels as they could (in the instructions before beginning the game), with the former being told additionally to attempt to achieve as high a rating as they could.

4 RESULTS

The challenge section of the HIT was completed by 288 players, 10 of whom failed to complete the survey. Thus, we ran our analyses on 278 players. Of these 278 players, 111 were randomly assigned into BLIND, 71 into CHOICE and 96 into RATINGS.

A summary of variable values is given in Table 1. Since the data was not normally distributed, we performed non-parametric tests for our analyses. First, we performed an omnibus Kruskal-Wallis test to look for significant differences across all three conditions. If found, we proceeded to perform three post-hoc Wilcoxon Rank-Sum tests with a Bonferroni correction to look for pairwise

Variable	Result
Challenge Time	$p = 0.003, H(2) = 11.60$
BLIND / RATINGS	$p = 0.026, W = 4197.5$
BLIND / CHOICE	$p = 0.010, W = 2927$
CHOICE / RATINGS	$n.s., W = 3565$
Levels Attempted	$p < 0.001, H(2) = 15.49$
BLIND / RATINGS	$p = 0.005, W = 3976$
BLIND / CHOICE	$p = 0.002, W = 2737$
CHOICE / RATINGS	$n.s., W = 3588$
Levels Completed	$p = 0.002, H(2) = 12.86$
BLIND / RATINGS	$p = 0.02, W = 4128.5$
BLIND / CHOICE	$p = 0.004, W = 2824.5$
CHOICE / RATINGS	$n.s., W = 3608.5$
Player Rating	$n.s., H(2) = 0.16$
Highest Level Rating	$n.s., H(2) = 3.67$
Interest/Enjoyment	$n.s., H(2) = 0.21$
Perceived Competence	$n.s., H(2) = 4.58$
Perceived Choice	$n.s., H(2) = 1.33$
Effort/Importance	$n.s., H(2) = 0.84$

Table 2: Summary table of statistical results from analysis. The first row for each variable is the omnibus Kruskal-Wallis test. Additional rows for a variable, if any, are the post-hoc Wilcoxon Rank-Sum tests with a Bonferroni correction for pairwise comparisons of the three experimental conditions.

significant differences between the conditions. A summary of all comparisons is provided in Table 2.

For *Challenge Time*, *Levels Attempted* and *Levels Completed*, we found significant differences across all three conditions. We found no pairwise difference between CHOICE and RATINGS but in both CHOICE and RATINGS, players spent significantly more time in the challenge levels and both attempted and completed significantly more number of levels than in BLIND.

We did not find significant differences across conditions for any of our other variables, be it *Player Rating* and *Highest Level Rating* (i.e. measures of engagement in terms of player and level ratings) or *Interest/Enjoyment*, *Perceived Competence*, *Perceived Choice* and *Effort/Importance* (i.e. self-reported measures of experience).

While offering players choice did not result in any observable impact in engagement or experience, we noticed an interesting pattern to the choices made by the player depending on the outcome of the previous level, as given in Table 3 ($\chi^2(4) = 37.3, p < .001$). Namely, following a win, players were more likely to select the recommended level over the easy one, while following a skip, players were more likely to select the easy level. As expected, the hard option was the least preferred choice in all conditions but it is interesting to note that while, both after a win and after a skip, the hard level was chosen about 11% of the time, after a loss, this percentage more than doubled to 23%.

Previous Result	Total	Easy	Rec.	Hard
Complete	592	40%	49%	11%
Forfeit	218	41%	36%	23%
Skip	170	57%	32%	11%

Table 3: Percentage of times each option was selected given previous match outcome ($\chi^2(4) = 37.3, p < .001$).

5 DISCUSSION

Based on our results, we conclude that *H1 is partially supported*. For all variables measuring the 'amount' of behavioral engagement, namely *Challenge Time*, *Levels Attempted* and *Levels Completed*, we observed significantly higher measures when informing players about their ratings and the underlying rating system than when not doing so. However such improvements were observed neither for the engagement metrics tied to the level and player ratings (i.e. *Player Rating* and *Highest Level Rating*), nor the self-reported measures of experience.

Additionally, *H2 has to be rejected*. The CHOICE condition did not produce any observable improvement in either engagement or player experience, as compared to the RATINGS condition.

These findings suggest that giving players feedback about their in-game performance and skill level, along with information about the degree of challenge posed by the levels in the game, and encouraging them to get a high rating, engages them to spend significantly more time playing the game and to utilize this time to attempt and complete a greater number of levels. Further, we can also conclude that players who were told about the working of the rating system did not exploit this additional information to attempt fewer levels and skip more levels so as to not risk lowering their rating. Rather, given that we observed no significant differences across conditions for *Player Rating*, we can say that exposing the players to the rating system made them play and finish more levels without worrying about their own rating and thereby playing defensively. It is worth mentioning that though both informed and uninformed players were told to complete as many challenge levels as they could in the game's initial instructions page, the informed players may have been further reinforced to do so by being encouraged to get as high a rating as they could when told about the ratings system, prior to the actual challenge levels.

Moving on to choice, our findings imply that allowing players the ability to choose the degree of difficulty of the levels impacted neither the engagement metrics nor the self-reported experience measures in any significant manner. The latter point is particularly interesting. Neither ratings feedback nor choice caused players to report increased values for any of the Intrinsic Motivation Inventory measures. For each measure, the median values obtained per condition as a percentage of the maximum possible value obtainable is given in Table 1. The maximum obtainable values for *Interest/Enjoyment*, *Perceived Competence*, *Perceived Choice* and *Effort/Importance* were 49, 42, 49 and 35 respectively. It is worth noting from this table that the measures for *Perceived Choice* are not significantly increased in the CHOICE condition as compared to the other two. This may be explained by the fact that while choices available

to players in the CHOICE condition more explicitly alter the levels faced by the player, players in the other two conditions still have the choice of skipping any level they want to. Based on the results, this may have been perceived as enough of a meaningful choice for most players. These findings are similar to those of Smeddinck et al. [35] who also observed that offering choice had no significant impact on the *Interest/Enjoyment* and *Effort/Importance* dimensions of the Intrinsic Motivation Inventory. Moreover, our findings related to choice are in line with Smeddinck et al.'s overall conclusion that variations in perceived autonomy do not greatly impact player enjoyment and motivation even though players prefer the presence of manual choices.

While choice failed to measurably improve player experience, it is interesting to note that the choice made by the player was often influenced by the outcome of the previous match, as shown in Table 3. Successfully completing a level seems to have not necessarily given players an increased sense of competence (also confirmed by our survey results). Rather, it made them less likely to choose the harder level and a completion was most likely to be followed by the player choosing the recommended level. This seems to suggest that after completing a level, and thus attaining a new, higher rating, the player usually gained enough confidence to not resort to selecting the easy level, but not enough confidence to select the hard level. Conversely, skipping a level was most likely followed by the player choosing the easy level. A possible explanation for this is a player might be more prone to skipping a level if that level is perceived to be too difficult to even attempt, and thus the player wanted the level immediately after that to be easier. Finally, and most interesting of all, the likelihood that the player selected the hard level doubled after a forfeit as compared to after a completion or a skip. A likely explanation for this is that after watching their rating go down due to a forfeit, players were more eager to get their rating back up again by completing a level with a higher rating than their own.

6 CONCLUSION

In this work, we explored the effects of giving feedback to players about their skill level by exposing them to the underlying rating system used to match them to levels of comparable difficulties. We also examined the effects of letting players choose the difficulty of levels served to them. We found that informing players about their own ratings, as well as the ratings of the levels they played, led to the players both attempting and completing an increased number of levels, as well as spending more time playing. Additionally, we found that offering players the ability to choose the difficulty of the levels did not have a significant impact on player engagement or experience, but that the choice of difficulty made by the player was often influenced by whether they won, lost or skipped the previous level.

Future work could examine if the findings presented in this paper also hold in other human computation games, particularly those in other problem domains. In addition to HCGs, similarly constrained games in other genres could also benefit from this work, namely, games which have procedurally generated or user generated content which is hard to reason about or manipulate in advance.

Furthermore, having established the demonstrably significant effects of ratings feedback, it is worth studying the effects of choice more thoroughly. We have already argued how the implicit choices present in the non-choice conditions of this game may have been as meaningful as the explicit choices presented in the choice condition, thus causing the latter to fail to bring about any significant improvement in players' sense of perceived choice. Designing and presenting players with more meaningful choices that allow for greater control over the flow of the game is something worth looking into in the future.

In a similar vein, most of the arguments presented to explain the pattern of player choice in the previous section are conjectures at best since we did not expect player choices to be influenced by previous match outcomes in this manner. More rigorously examining the impact of outcome history on player choice is fertile ground for future research, particularly having a condition that offers players choice over level difficulty without giving them any information about their own skill or exposing them to the underlying rating system.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under grant no. 1652537. We would like to thank the players, and the University of Washington's Center for Game Science for initial *Paradox* development.

REFERENCES

[1] Justin T. Alexander, John Sear, and Andreas Oikonomou. 2013. An investigation of the effects of game difficulty on player enjoyment. *Entertainment Computing* 4, 1 (Feb. 2013), 53–62.

[2] Alexander Baldwin, Daniel Johnson, Peta Wyeth, and Penny Sweetser. 2013. A framework of Dynamic Difficulty Adjustment in competitive multiplayer video games. In *Proceedings of the 2013 IEEE International Games Innovation Conference*. 16–19.

[3] Alexander Baldwin, Daniel Johnson, and Peta A. Wyeth. 2014. The effect of multiplayer Dynamic Difficulty Adjustment on the player experience of video games. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. ACM, New York, NY, USA, 1489–1494.

[4] Paul Black and Dylan Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice* 5, 1 (March 1998), 7–74.

[5] Jared E. Cechanowicz, Carl Gutwin, Scott Bateman, Regan Mandryk, and Ian Stavness. 2014. Improving player balancing in racing games. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play (CHI PLAY '14)*. ACM, New York, NY, USA, 47–56.

[6] Kate Compton, Heather Logas, Joseph C. Osborn, Chandranil Chakrabortti, Kelsey Coffman, Daniel Fava, Dylan Lederle-Ensign, Zhongpeng Lin, Jo Mazeika, Afshin Mobramaein, Johnathan Pagnutti, Husacar Sanchez, Jim Whitehead, and Brenda Laurel. 2016. Design lessons from Binary Fission: a crowd sourced game for precondition discovery. In *Proceedings of the 1st International Joint Conference of DiGRA and FDG*.

[7] Seth Cooper, Sebastian Deterding, and Theo Tsapakos. 2016. Player rating systems for balancing human computation games: testing the effect of bipartiteness. In *Proceedings of the 1st International Joint Conference of DiGRA and FDG*.

[8] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit Players. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (Aug. 2010), 756–760.

[9] Valve Corporation. 2008. *Left 4 Dead.* Game. (2008).

[10] Mihaly Csikszentmihalyi. 1990. *Flow: the psychology of optimal experience.* Harper and Row, New York.

[11] Drew Dean, Sean Gaurino, Leonard Eusebi, Andrew Keplinger, Tim Pavlik, Ronald Watro, Aaron Cammarata, John Murray, Kelly McLaughlin, John Cheng, and Thomas Maddern. 2015. Lessons learned in game development for crowdsourced software formal verification. In *Proceedings of the 2015 USENIX Summit on Gaming, Games, and Gamification in Security Education*. USENIX Association, Washington, D.C.

[12] Alena Denisova and Paul Cairns. 2015. Adaptation in digital games: the effect of challenge adjustment on player performance and experience. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '15)*. ACM, New York, NY, USA, 97–101.

[13] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 1013–1022.

[14] Arpad E. Elo. 1978. *The rating of chessplayers, past and present*. Arco.

[15] Stefan Engeser and Falko Rheinberg. 2008. Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion* 32, 3 (Sept. 2008), 158–172.

[16] Clive J. Fullagar, Patrick A. Knight, and Heather S. Sovern. 2013. Challenge/skill balance, flow, and performance anxiety. *Applied Psychology* 62, 2 (April 2013), 236–259.

[17] Mark E. Glickman. 1999. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48, 3 (1999), 377–394.

[18] Mark E. Glickman. 2001. Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics* 28, 6 (Aug. 2001), 673–689.

[19] Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkill(TM): a Bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*. MIT Press, 569–576.

[20] Robin Hunicke. 2005. The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology (ACE '05)*. ACM, New York, NY, USA, 429–433.

[21] Martin Jennings-Teats, Gillian Smith, and Noah Wardrip-Fruin. 2010. Polymorph: dynamic difficulty adjustment through level generation. In *Proceedings of the 2010 Workshop on Procedural Content Generation in Games (PCGames '10)*. ACM, New York, NY, USA, 11:1–11:4.

[22] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker motivation in crowdsourcing - a study on Mechanical Turk. In *Proceedings of the Americas Conference on Information Systems*.

[23] Mohammad M. Khajah, Brett D. Roads, Robert V. Lindsey, Yun-En Liu, and Michael C. Mozer. 2016. Designing engaging games using Bayesian optimization. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5571–5582.

[24] Christoph Klimmt, Tilo Hartmann, and Andreas Frey. 2007. Effectance and control as determinants of video game enjoyment. *Cyberpsychology & Behavior* 10, 6 (Dec. 2007), 845–847.

[25] Kai Kuikkaniemi, Toni Laitinen, Marko Turpeinen, Timo Saari, Ilkka Kosunen, and Niklas Ravaja. 2010. The influence of implicit and explicit biofeedback in first-person shooter games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Atlanta, GA, 859–868.

[26] Derek Lomas, Kishan Patel, Jodi L. Forlizzi, and Kenneth R. Koedinger. 2013. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, Paris, France, 89–98.

[27] J. Derek Lomas, Ken Koedinger, Nirmal Patel, Sharan Shodhan, Nikhil Poonwala, and Jodi L. Forlizzi. 2017. Is Difficulty Overrated?: The Effects of Choice, Novelty and Suspense on Intrinsic Motivation in Educational Games. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Denver, CO, 1028 – 1039.

[28] Winter Mason and Duncan J. Watts. 2009. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '09)*. ACM, Paris, France, 77–85.

[29] Elisa D. Mekler, Florian Bruhlmann, Alexander N. Tuch, and Klaus Opwis. 2017. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior* 71, C (June 2017), 525–534.

[30] Richard M. Ryan and Edward L. Deci. 2000. Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemporary Educational Psychology* 25, 1 (Jan. 2000), 54–67.

[31] Richard M. Ryan and Edward L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* 55, 1 (2000), 68–78.

[32] Anurag Sarkar, Michael Williams, Sebastian Deterding, and Seth Cooper. 2017. Engagement Effects of Player Rating System-Based Matchmaking for Level Ordering in Human Computation Games. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*. Hyannis, MA.

[33] Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research* 78, 1 (March 2008), 153–189.

[34] Kristin Siu and Mark O. Riedl. 2016. Reward systems in human computation games. In *Proceedings of the SIGCHI Annual Symposium on Computer-Human Interaction in Play*.

[35] Jan D. Smeddinck, Regan L. Mandryk, Max V. Birk, Kathrin M. Gerling, Dietrich Barsilowski, and Rainer Malaka. 2016. How to present game difficulty choices?: exploring the impact on player experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5595–5607.

[36] Alex Cho Snyder and Mario Izquierdo. 2014. *Jumpcraft*. Game [PC]. (2014).

[37] MADDALENA TARAS. 2003. To feedback or not to feedback in student self-assessment. *Assessment & Evaluation in Higher Education* 28, 5 (Oct. 2003), 549–565.

[38] Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Vienna, Austria, 319–326.

[39] Michael Williams, Anurag Sarkar, and Seth Cooper. 2017. Predicting Human Computation Game Scores with Player Rating Systems. In *Predicting Human Computation Game Scores with Player Rating Systems. In: Munekata N., Kunita I., Hoshino J. (eds) Entertainment Computing – ICEC 2017. ICEC 2017.*