Effect of Timer, Top Score and Leaderboard on Performance and Motivation in a Human Computing Game

Anjum Matin
Mardel Maduro
Rogerio de Leon Pereira
Olivier Tremblay-Savard
tremblao@cs.umanitoba.ca
University of Manitoba
Department of Computer Science
Winnipeg, Manitoba, Canada

ABSTRACT

The development of human computing games requires the implementation of different game mechanics to make them challenging, interesting and motivating. These mechanics are often borrowed from popular video games, but their outcomes on the quality of solutions obtained and player motivation are not fully understood. We analyze the effect of showing a timer, an achievable (top) score and a live leaderboard on players' scores, puzzle completion time and motivation using different versions of a human computing game. We show that presenting a top score on a puzzle results in better solutions, but at the expense of completion time, whereas the presence of a timer has the opposite outcome. As for the live leaderboard, we have observed an almost significant interaction effect with the timer. This work offers guidance for human computing game developers about what to expect from these different game mechanics, and how players react to them.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; • **Applied computing** → **Computer games**.

KEYWORDS

Gamification, Crowdsourcing, Interaction Design, Human Computation, Human Computing Game

ACM Reference Format:
Anjum Matin, Mardel Maduro, Rogerio de Leon Pereira, and Olivier Tremblay-Savard. 2020. Effect of Timer, Top Score and Leaderboard on Performance and Motivation in a Human Computing Game. In *International Conference on the Foundations of Digital Games (FDG '20), September 15–18, 2020, Bugibba, Malta.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3402942.3403000

1 INTRODUCTION

Human computing (process of recruiting humans to help solve problems that are complex for computers), *crowdsourcing* (practice of obtaining information from a large crowd of people, typically online) and *citizen science* (recruitment of the general public to help collect and analyze scientific data) have developed rapidly in the last decade. Those practices are now being considered as valid options to solve various types of problems which can benefit from human abilities. Platforms such as Amazon Mechanical Turk [20] or Microworkers [45] have been created for the distribution of crowdsourcing tasks, where the tasks are completed by human "workers" in exchange of a monetary compensation. There is also an increasingly high number of human computing, crowdsourcing and citizen science games being deployed, which aim to solve problems from many different areas, such as molecular biology [11, 22, 26], health [1], ecology [5, 6], neuroscience [23], astronomy [3, 7], quantum physics [37] and deep learning [38]. Participants playing these games generally do it for free, which can be an advantage for requesters who need to distribute a large amount of *human-intelligence tasks*. However, there is a major trade-off: the game needs to be interesting and motivating enough to attract and retain a large enough player-base, and finding the right balance between accurate data collection and a fun gaming experience is an extremely difficult task [34].

One way to make human computing games more captivating is for them to emulate commercial video games by incorporating several game mechanics, such as scoring systems, timers and leaderboards. However, it is not clear how efficient these game features are at keeping the players interested and motivated in the game and if they can potentially interfere with the quality of the solutions produced. This is why research in human computing games is now focusing more and more on the analysis of different game features and game mechanics [33–36, 39]. The objective is to learn more about which game mechanics work well in the context of these *games with a purpose*, and how to implement them successfully.

In this paper, we present a study of three game mechanics in the context of a human computing game in genomics: a timer (setting a time limit on a puzzle), a top score (representing a score that is achievable), and a live leaderboard (points are rewarded for completing puzzles, and the player's progress is shown on a leaderboard that is always present on the screen). Using different versions of the same human computing game, we evaluate the effect of those three game elements on the players' scores, puzzle completion times and motivation in the game. Moreover, we study how all these mechanics interact together by testing all the possible combinations of them being present or not.

FDG '20, September 15–18, 2020, Bugibba, Malta

ACM ISBN 978-1-4503-8807-8/20/09…$15.00
https://doi.org/10.1145/3402942.3403000

2 RELATED WORK

Several recent studies have been devoted to the analysis of game mechanics in human computing games. Siu *et al.* [35] have investigated competitive and collaborative scoring mechanisms in a human computing game and found that competitive scoring is better at engaging players than collaborative scoring, while yielding results that are just as accurate. Siu and Riedl [34] have observed that offering a choice of rewards to players improves task completion and player experience. Another study on a *Super Mario Bros* based human computing game has shown that players in the collaborative multiplayer version produced better solutions than players in the competitive multiplayer version, although the latter was found to be more challenging [33]. Tremblay-Savard *et al.* [39] have studied how a market system can be used in a collaborative human computing game to promote cooperation, and how a skill and a challenge system can guide players into accomplishing actions that are beneficial to the collaborative system.

To the best of our knowledge, only Phylo [22] and FoldIt [17] have implemented a game mechanic that gives the player some idea of what an achievable score (or number of moves) can be on any given level. In Phylo, players have to produce an alignment of sequences that is at least as good as a machine-computed alignment to proceed to the next level. This score that must be attained is called the "par" [22] and it is guaranteed to be achievable. However, the "par" score is not always a good representation of the best score that can be obtained by a player, which results in some puzzles being harder than others simply because they offer a "lower possibility of improvement" [31] (e.g. when the machine-computed alignment is reasonably good, the par is more difficult to reach). The current version of Phylo also shows the top score [15], which corresponds to the best score achieved by a player on a specific level, although the authors have not studied the effect of presenting it to the players. FoldIt ran an online experiment on a three-star system for their tutorial levels [17]. Instead of being linked to the score of a folded protein, the star system was linked to the total number of moves used by the players to complete each level. Players would get three stars if they completed the level using an "ideal" number of moves, and lost stars for using additional moves. The authors found that the three-star system resulted in players taking more time per move, producing solutions with fewer moves and retrying the levels more often. However, they did not observe any significant benefit on the retention of players.

Timers are often used in cooperative human computing games, such as the ESP game [42], Peekaboom [43], TagATune [25] and Ask'nSeek [28], to assure that both cooperating players complete the levels in a reasonable time. Timers have also been used in the context of collaborative leisure, such as the Story Mashup game [29, 40], to make sure that "players cannot stall the game dynamics" [40]. In single-player human computing games, timers are used mostly as an attempt to make them more interesting and challenging. For example, a timer was initially integrated in the first version of Phylo as one of the mechanisms designed to "increase the entertaining value of the game" [22], although it has been retired from the current online version [15]. Lomas *et al.* [27] have studied the effect of giving players different time limits to complete levels in an online *Battleship-style* game where players must estimate numbers on a number line to sink enemy ships. They observed that giving more time to the players increased success (*i.e.* accurate estimates) and engagement (measured as a combination of time spent in the game and number of levels attempted), but noted that the online nature of the experiment did not allow to gain much qualitative insight into why the players responded in this manner. Denisova and Cairns [14] explored the effect of dynamic difficulty adjustment on player experience and performance by manipulating the speed of a timer to make a shooting game more or less challenging depending on player performance. They observed that players in the dynamically-adjusted timer condition felt more immersed in the game and obtained scores that were less variable than the players in the control condition with a normal timer.

Leaderboards are one of the most common features of human computing games, as they are often used as a gamification technique to stimulate the more competitive participants and encourage them to perform better and eventually get to the top of the leaderboard [18, 21, 24]. Several studies have shown that different personality types (extroverted vs introverted for example) respond differently to leaderboards and have called for more carefully designed and personalized leaderboards [10, 16, 21]. Bowey *et al.* [4] have shown that manipulating a leaderboard to simulate success increases the player's perception of competence, autonomy, presence, enjoyment and positive affect. The authors have also noted that it would be interesting to see if it would be even more effective to display a leaderboard in real-time, instead of only presenting the leaderboard at the end of the game [4]. While in most human computing games the leaderboards are only accessible in a menu or shown after completing a level [8, 11, 12, 19, 23, 37, 39, 43, 44], there are relatively few examples of live leaderboards, or in other words, leaderboards that are continuously present on the screen (TagATune being one of them [25]).

3 SYSTEM OVERVIEW

3.1 Description of the human computing game

We have developed a web-based casual puzzle game to solve a difficult problem in genomics: finding the shortest sequence of biological operations that can transform one ordered sequence of genes into another. This problem is NP-hard in the presence of duplicate genes [9], and it was targeted to eventually devise new heuristics from the players' solutions (we refer the reader to [32] for more details on the development of this game and a platform of games for genome analysis).

In the game, the sequences of genes are simply represented as sequences of colored shapes (see Figure 1). We used seven distinct shapes and three different colors (safe for color blindness) to represent the different genes in the sequences. This provides enough gene representations to study specific groups of genes, like tRNA genes for example, as there are 20 possible types of these genes.

Players are shown a mutable sequence of colored shapes that they must transform into a target sequence. Each move (i.e. operation) applied by the player increases the score by 1, and similarly to golf, the objective is to complete each puzzle with the minimum score. The operations that are available to the players correspond to biologically relevant evolutionary events: duplications (copying selected colored shapes to another location), deletions (removing

Figure 1: The game interface for the timer, top score and leaderboard condition. The interface can be separated into three panels: panel A (green box) is the live leaderboard, panel B (red box) is the player information panel, and panel C (yellow box) is the game panel.

the selected colored shapes) and inversions (inverting the order of the selected colored shapes). There is an additional constraint on inversions events, which is motivated by biology: the inversions must occur around a "pivot" in the sequence of genes, represented as a black dot near the center of the sequence (this corresponds to the *terminus of replication* in bacterial genomes). In other words, this means that inversion operations must be applied to a selection of colored shapes that includes the black dot to be valid.

In a normal game setting, the ordered sequences of genes that constitute the puzzles can easily be extracted from public databases (e.g. GenBank [2]) for any pairs of genomes of interest. For this specific experiment though, synthetic datasets were generated to have a better control over the puzzle difficulty (see section 5.2 for a description of the synthetic datasets and how they were produced).

3.2 Implementation of the three game mechanics

Timer: To assure that a reasonable time limit was set for each level, three lab members with varying degrees of experience with the game played all the levels of all datasets (see section 5.2 for a full description of the datasets), and their average completion time was calculated for each level. The time limit on each level was set to the calculated average time to which 45 to 90 seconds were added, depending on the level's difficulty (*i.e.* +45s for levels 1-2, +60s for levels 3-4, + 75s for level 5 and +90s for level 6). As explained in section 5.2, each dataset has 6 levels which gradually increase in difficulty, which is why more time is added as the levels go up. The idea was to strike a balance between giving a reasonable amount of time to complete the levels, and creating a challenge with the timer. The timer starts counting down as soon as the level is loaded and if it is incomplete when all the time has expired, the participant is prompted to restart the level (skipping the level was not an option).

Top score: The top score aims to represent the best (in this context, lowest) score achieved by a player on the level. In this study, we set the top score for each level to the number of events that were simulated plus 1 (see section 5.2). This was done to simulate a real top score, which would leave some room for improvement. Note that the number of events simulated is not necessarily the lowest number of events required to solve the puzzle, as some shortcuts can be found by combining events together (e.g. two consecutive deletions that were simulated could be resolved by one deletion of a block of consecutive genes).

Leaderboard: The leaderboard was populated with 9 simulated players in addition to the current player. Points were assigned to the simulated players in such a way that each real participant could potentially reach the top of the leaderboard during a game session. The point system for the leaderboard assigns 50 points for completing a level, and additional points depending on the presence/absence of the timer and the top score. Table 1 presents the distribution of points for each of the possible game conditions.

Table 1: Distribution of points per puzzle for each condition when the leaderboard is present. 'T' represents timer and 'TS' represents top score, 'time' represents the remaining time in seconds, and 'score' represents the player's score.

Conditions	Calculation of points
T:OFF, TS:OFF	50
T:ON, TS:OFF	50 + 1 x time
T:OFF, TS:ON	50 + max{50 x (TS - score +1), 0}
T:ON, TS:ON	50 + 1 x time + max{50 x (TS - score + 1), 0}

3.3 Game interface

The game was built in Unity 3D using C#. As shown in Figure 1, the game interface can be divided into three parts: the live leaderboard (A), the player information panel (B) and the game panel (C).

A: Leaderboard
This panel contains a live leaderboard showing the standings of the players according to their points. After completing each level, the players are rewarded with points and their ranking is updated. The top three positions have a gold, silver and bronze medal associated with them (similarly to the colors that were used in [4]) and the current position of the player is highlighted in light blue.

B: Player information panel
There are five components in the player information panel. On the left-hand side, there is a pause button. Upon clicking on the pause button, a pause menu will show up and from there the player can go to main menu, restart the level or resume the game. The current level number is displayed next to the pause button. The timer appears in the center. On the right-hand side of the screen, the current score (number of moves played on the level) and top score (under the "best" label) are displayed.

C: Game panel
In the game panel, the puzzle is shown at the top: the first row is the *target* sequence (not modifiable) and the second row is the *mutable* sequence. The players need to transform the mutable sequence into the target sequence by performing a series of possible operations, which can be applied by first selecting colored shapes and clicking on one of the buttons at the bottom of the panel: dup (duplication), del (deletion) or inv (inversion). Once the two rows of sequences are matched completely, the "Level completed" panel will appear and will prompt the players either to go to the next level or restart the current level to achieve a better score.

4 RESEARCH QUESTIONS

Considering the two main dependent variables, which are the score obtained by a player and the completion time for each level (puzzle), and the player motivation (as determined by qualitative analysis), we aim to answer these research questions:

RQ1: *What is the effect of setting a time-limit for each puzzle, using a timer?*
RQ2: *What is the effect of showing the top score?*
RQ3: *What is the effect of showing the live leaderboard coupled with a context-dependent point system?*
RQ4: *How will these three different game mechanics (timer, top score and live leaderboard + point system) interact together?*

5 EXPERIMENTS

5.1 Experimental Design

We used a 2^3 mixed factorial design, with two within-subjects factors (top score, leaderboard) and one between-subject factor (timer) which can all be either on or off. The timer was chosen to be a between-subject factor to avoid a potentially pervasive effect where players being introduced to a timer would think that they always need to complete levels as quickly as possible, even when it is not there anymore.

24 participants were recruited in total to play the game individually. Half of them played with the timer present all the time, and the other half without it. Every participant played all the four different combinations (on/off) of the top score and live leaderboard.

Every participant answered a short demographic questionnaire and completed an in-game tutorial before starting the experiment. Each game session (one for each condition) contained six different levels of increasing difficulty (see section 5.2), and the first level was used for training and briefly explaining the new game condition, thus allowing the players to get accustomed to it. The four conditions were played in different orders by the participants to avoid sequence effects. Moreover, the four datasets were shuffled between participants to ensure that the same dataset is not always associated with the same game condition.

After playing each game session, a short interview was conducted to get the player's opinion on the last game condition played. A longer interview was conducted at the very end of the experiment to get the player's overall opinion on the game, the different conditions and the three game mechanics studied (in this game specifically and also other games in general). For the qualitative data analysis, interviews were transcribed by one of the authors and analyzed by all the authors. Players' performance in the game (time to complete a level and scores) was recorded for all the players, for each level of each condition.

5.2 Generation of the datasets

We used synthetic datasets for this experiment, created using a level generator which simply creates an original sequence (the target) and modifies it by applying random events to obtain the mutable sequence. We created four distinct datasets consisting of 6 levels of slightly increasing difficulty. To increase the difficulty, we generated longer sequences of colored shapes (from 10 to 18) and simulated more events (from 3 to 6) between the target and the mutable sequence. Each corresponding level in each of the four datasets was simulated using the same parameters, to assure a similar level of difficulty between datasets. The goal of slowly increasing the difficulty was to alleviate the potential consequences on the results of having only one difficulty level that would be either too easy or too hard, and to keep the players challenged as they

progress (i.e. keep them in the so-called *flow channel*, in between boredom and anxiety [30]).

Note that estimating or predicting the difficulty of a puzzle is not simple and often requires deeper analyses (see [31, 41] for examples). Based on prior testing of our game, we feel confident that most of the difficulty is coming from the size of the sequences and the number of simulated events, but there could definitely be other factors at play. For instance, it is possible that some specific sequences of events could be harder to find than others.

5.3 Participant demographics

24 participants (16 male, 8 female) were recruited using posters that were placed in several buildings of the authors' university for convenience. Participants were compensated with a $10 gift card for their time commitment (45 minutes on average). The participants' average age was 25.8 and their average time playing video games was 5.7 hours/week. All the participants were university students and most of them were at the undergraduate level.

6 RESULTS

6.1 Quantitative results

Average scores. The average scores per level for each game condition are shown graphically in Figure 2. Recall that the score represents the total number of moves made to successfully complete a level, hence lower scores represent better performances.

Clearly, mean scores are lower when the timer is off, and the lowest mean (5.77) is obtained in the *top score on / leaderboard on* condition. This can be explained by the fact that the top score gives players an idea of what an achievable score can be, and the combination of the point system with the leaderboard further encourages players to achieve a low score (better performance). On the other hand, the highest mean score (7.47) occurred when the timer was on, in the *top score off / leaderboard on* condition. Players trying to complete the level faster to get more points, coupled with the absence of the top score in this condition, probably encouraged the players to complete the levels as quickly as possible without as much consideration for the number of moves they were using.

Average level completion times. The average completion times per level for each game condition are shown graphically in Figure 3. Note that since players can restart a level (or are sometimes forced to restart a level if the time runs out in the *timer on* condition), we recorded the full time that was necessary to complete the level (*i.e.* considering all the attempts).

We can clearly observe that the average completion times were considerably lower whenever the timer was present. Interestingly, the total number of restarts for the players in the *timer on* group was 92, compared to only 19 in the *timer off* group. In other words, even though the participants playing with the timer had to restart the levels a lot more often, the total level completion times of all these attempts were still lower on average that those of the *no timer* group.

The variability of the completion times is also much higher when the timer is absent, as demonstrated by the standard deviations which are approximately twice as large compared with the ones of the *timer on* group. Mean completion times are very similar between

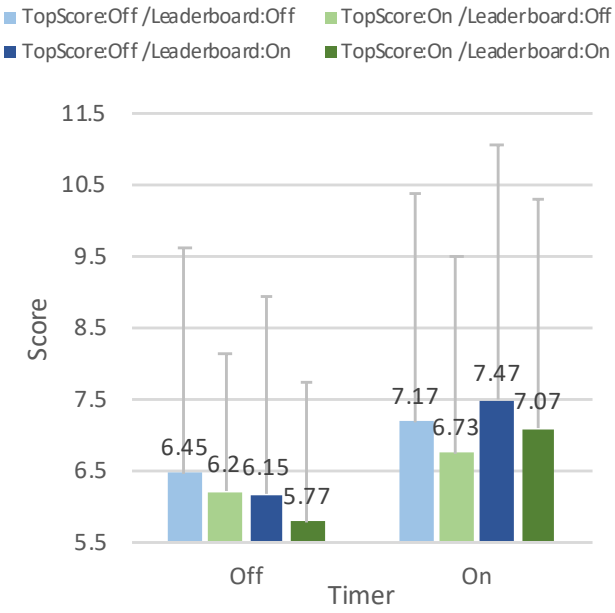

Figure 2: Mean scores and standard deviations according to the presence of all three independent variables (timer, top score and leaderboard). Blue colors represent TopScore:Off whereas green colors represent TopScore:On. Light colors represent Leaderboard:Off whereas dark colors represent Leaderboard:On. Recall that lower scores represent better performances.

all four possible conditions when the timer is present. However, the completion times are significantly higher when the top score is on and the timer is off, which indicates that players are taking more time in those conditions as they try to reach or beat the top score. We also calculated how much the solutions can be improved (in terms of score) per amount of extra "human-computation" time when the timer is off. We computed that we obtain a score that is 0.69 lower on average for each extra 10 seconds of play (in the *top score on, timer off* condition, vs the *timer on, top score off* condition).

Main effects and interactions. A mixed factorial ANOVA was performed to check for main effects and interactions of the three factors studied (timer, top score and leaderboard) on players scores and completion time (see Table 2). The analysis detected a main effect of the timer and the top score on both the scores and completion times. However, no significant main effect of the leaderboard was detected on the two dependent variables.

Interestingly, we did not observe significant interactions between the different mechanics, only an almost significant interaction ($F = 3.27$, $p = 0.07$, see Table 2) between the timer and the leaderboard, reinforced by the point system. We looked into the simple effects of the timer on both the score and completion time (see Table 3). It turns out that the timer has a significant effect on the score only when the leaderboard is present, whereas the timer has a significant effect on each of the four conditions for the completion time. This

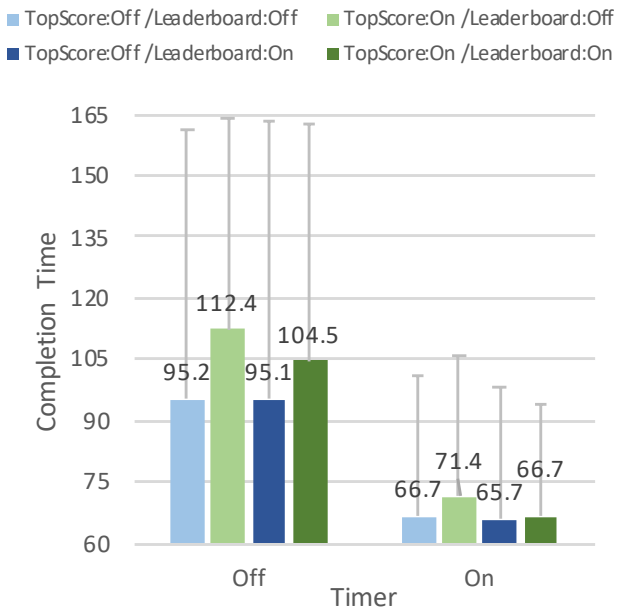

Figure 3: Mean level completion times (in seconds) and standard deviations according to the presence of all three conditions (timer, top score and leaderboard). Blue colors represent TopScore:Off whereas green colors represent TopScore:On. Light colors represent Leaderboard:Off whereas dark colors represent Leaderboard:On.

Table 2: Mixed factorial ANOVA results for the score and completion time dependent variables. The main effects of the timer (between-subject factor), the top score (within-subject factor) and leaderboard (within-subject factor), and their main interaction effects are shown. Significant codes (Sphericity assumed): p<0.05 '*'.

Factors	Score		Completion Time	
	F value	p value	F value	p value
Timer (Between)	5.121	**0.025***	28.622	**0.000***
Top score (Within)	4.843	**0.03***	5.869	**0.017***
Leaderboard (Within)	0.08	0.895	1.075	0.302
Top score * Timer	0.90	0.765	2.481	0.118
Leaderboard * Timer	3.27	0.07	0.031	0.861
Top score * Leaderboard	0.021	0.886	0.483	0.488
Top score * Leaderboard * Timer	0.058	0.811	0.064	0.800

interaction can be observed on Figure 2, where we can see that the lowest average scores (better scores) are obtained when the timer is off and the leaderboard is on. On the other hand, some of the highest average scores (worst scores) are obtained when both the timer and the leaderboard are on. In other words, the point system that comes with the leaderboard seems to encourage the players to make more moves in the presence of the timer, which is probably a consequence of the players just trying to complete the level as fast as possible.

Table 3: Simple effects of Timer (on vs off) for each of the four combinations of the other two factors (top score (TS) and leaderboard (LB)). Significant codes: p<0.05 '*'.

Conditions	Score		Completion Time	
	F value	p value	F value	p value
1) TS:ON, LB:ON	7.221	**0.008***	18.948	**0.000***
2) TS:ON, LB:OFF	1.509	0.2220	26.319	**0.000***
3) TS:OFF, LB:ON	5.044	**0.0270***	9.046	**0.003***
4) TS:OFF, LB:OFF	1.521	0.2200	9.579	**0.002***

6.2 Qualitative results

Favorite game session. The participants were assigned to either the timer or no timer experiment, so they had in common only the presence or absence of top scores and leaderboards. When asked about their preferred combination of game mechanics, 46% of the participants preferred the condition with both the leaderboard and the top score, followed by 31% for the top score without leaderboard and 23% for only the leaderboard. None of the participants voted for the version in which both game mechanics (top score and leaderboard) were missing. In other words, the majority of the participants (46% + 31% = 77%) preferred to see the top score, either with or without the leaderboard.

Opinions on the top score. More than half of the participants mentioned that they were paying close attention to the top score, that it forced them to think more about their choices of moves, and that seeing the top score made the game more challenging and interesting. Some of them also mentioned that they did not appreciate losing the top score after playing a session where it was present.

A few participants pointed out that the top score was giving them a sense of direction, allowing them to know how well they were performing in the game:

> *The top score is interesting, because it gives you an idea of the optimal solution. (P13)*

These kinds of responses from the participants are in line with our observation of better (lower) scores in the presence of this game feature (see the green bars in Figure 2), in the sense that we know that they were clearly paying attention to it and the added challenge of reaching or beating that score seemed to be motivating players to find better solutions.

On the other side of the coin, some participants alluded to the potential for the top score to cause anger and discouragement:

> *In any game with the top score, you know there is always somebody who can do crazy things, which most people can't do. (P3)*
> *Up to some point it's okay if you can beat it, otherwise*

it's frustrating. (P14)
I feel frustrated when I can't reach the top score. (P24)

Opinions on the live leaderboard. Ten different participants mentioned that they enjoyed seeing their position in the leaderboard, how they ranked among other players, seeing their names go up in the live leaderboard and the rewarding aspect of getting points based on the other game mechanics present.

A participant mentioned that climbing up in the leaderboard was a powerful motivation to perform better in the game:

I think I tried harder when the leaderboard was on. (P20)

Other participants pointed out that they did not feel the same sense of progression and improvement when the live leaderboard was removed in a different game condition:

No motivation because no leaderboard, no progress. (P4)
I was trying to improve but I couldn't compare with others now. (P10)

A few other participants were uninterested by the live leaderboard and did not see the need for having it on the screen at all times, like P6:

I didn't pay attention to the leaderboard. It's good but you don't need it there all the time. (P6)

Interestingly, another participant suggested that the live leaderboard could be distracting for the users, which could especially be true in the context of a larger player-base, where names would be constantly moving up and down in the updated rankings:

It disrupted my concentration. (P23)

Opinions on the timer. Participants who played in the *timer on* group felt challenged to complete the level within a given time limit and to avoid a forced restart. It kept their completion times low (see Figure 3), but at the expense of a higher number of moves as we have seen previously (see Figure 2). Keeping track of the timer and the top score at the same time was challenging for some players:

Dealing with the top score and the timer was the most challenging part in my last game session. (P2)
Keeping pace with time left is hard, [the game] actually requires time to think. (P14)

Nevertheless, several players seemed to enjoy the additional challenge that the timer brought to the game. The timer exercises a certain kind of pressure that helps some players to keep focused on the game and some participants appreciated the thrill of quickly coming up with a solution:

Timer is good, it adds more challenge, makes it more interesting. (P16)
It keeps you more focused. (P18)
It's good, stressful in a good way. (P6)
It's good, because you are thinking of different strategies in a short amount of time and without it, I would get bored. (P24)

Interestingly, some players disliked the timer for the exact same reason of the added pressure. Some players pointed out that for them, a puzzle game is supposed to be relaxing, and the timer is creating the opposite effect:

I didn't like the timer, it added pressure. (P12)
I like the idea of giving a time limit, but for a relaxation game you actually don't need a timer. (P4)
I don't like a timer for a puzzle game because it is meant for relaxation. (P20)

7 DISCUSSION

In our game, similarly to Phylo, the ideal number of moves was presented to the players as the top score (just a number), whereas the three-star system had a more elaborate representation, using a test tube filled with liquid that would drain every time a move was made, and the stars were overlaid on top of the tube. As mentioned in the related work section, Phylo's top score and its effect on players was not analyzed by the developers, so a comparison can only be made with FoldIt. Our results about the effect of showing a top score are in line with some of those of the three-star system in FoldIt [17]: giving the players an idea of an "ideal" number of moves influences the players to take their time and produce solutions with fewer moves. Our study suggests that the same effect on the number of moves can be attained, in the context of a different game, without the extra visual representation (e.g. of the test tube emptying). However, as opposed to the three-star system, neither the top score alone nor the top score coupled with the live leaderboard had an effect on the number of level restarts (only the timer had a significant effect on restarts). In other words, players were more likely to restart a level because they ran out of time.

One important detail about the game is that suboptimal solutions to the puzzles are easier to find than optimal solutions. When the timer is present (and even more so when the leaderboard is present as well), players seem to take the easier route to solve the puzzles quicker, but using more moves. This is what could explain the almost significant interaction between the timer and the leaderboard. It is possible that either the point system was too generous for solving puzzles quickly (or not generous enough for beating the top score), or that it was just a lot easier to get points by beating a level quickly rather than trying to find a short sequence of optimal moves. A similar observation was made in [39], where the challenges (actions that could be completed for in-game rewards) implemented in the game were ignored by the players if they were too hard. In the *top score on, timer on and leaderboard on* condition, perhaps a stronger reward (in terms of points) for reaching the top score could have tipped the scales, but there is a chance that it would just be too hard for the players to optimize both objective functions.

Although [27] did not test their battleship-like game without a timer (the timer was always there, but the authors tried different time limits), our results are in line with theirs: longer time limits (in our case, no timer is equivalent to an infinite amount of time) increase the players' performance. The authors of [27] also found that longer time limits increase engagement (measured as a combination of number of trials (*i.e.* games) and total time played), but the authors mentioned that their online study with anonymous participants did not allow them to get "insight into why subjects are responding the way that they do".

Although our study did not allow us to directly measure player retention (game sessions were the same length for all participants),

our qualitative analysis highlighted a lot of ambivalence from our pool of participants regarding the different game elements. Some players liked the added challenges and excitement provided by the timer, the top score and the leaderboard, while others disliked the added pressure and wanted a more relaxing game experience. The only thing that was constant across all of the participants was that they preferred to have at least one of these game elements present. In terms of the presence or absence of the timer, some players might enjoy more the challenge of slowly finding the shortest sequence of moves, whereas others might be motivated more by the challenge of quickly finding a solution before the time runs out, which can be viewed as two completely different types of puzzles.

Our decision to test a live leaderboard (instead of a "regular" leaderboard, showing up only occasionally or when the user wants to) was based on a suggestion in [4] that it might be more effective in promoting feelings of competence and enjoyment. Although we did not compare live leaderboards with "regular" leaderboards (only live leaderboard on/off), it was not obvious from our results that having the leaderboard present on the screen at all times was beneficial, with some players even mentioning that they did not pay attention to it at all.

Based on our study, some lessons can be drawn for the design of human computing games. Firstly, from a player motivation perspective, it might be beneficial to implement multiple game features and automatically alternate between the different "game modes" to keep different types of players motivated. Another option could be to let the players choose which game mode they want to play. Since the authors of [34] have shown that letting the players choose their rewards improved task completion and player experience, it is possible that letting the players choose their favorite game mode could have similar beneficial effects. Secondly, from a data acquisition point of view, timers should be implemented when a large number of solutions are necessary in a short amount of time, but they should be avoided when the task is very hard or when top quality solutions are required. Thirdly, presenting a top score is beneficial to guide the players toward higher quality solutions, but it can also be frustrating if it is too hard to reach. To mitigate this, human computing game designers should consider allowing the players to advance to the next level even if the top score was not reached and/or retiring puzzles when the top score seems unbeatable (e.g. when it has not been improved upon in the last x player attempts). Fourthly, a points system coupled with a (live) leaderboard can be motivating for some (but not all) players, but serious thought has to be put in its design. Building a balanced point system (rewarding players proportionally to the complexity of the task) is a difficult but necessary exercise, since players will always find the easiest route to get more rewards. Moreover, although we did not test it, top leaderboard rankings might quickly become unreachable in a real setting, which could be demotivating for newer players. We suggest that implementing several different leaderboards (e.g. daily, weekly, monthly, all time, etc.) could be beneficial for player motivation. Once again, customizability might be the key to player motivation: some players might be motivated by a live leaderboard (of their choice, e.g. daily or monthly), while others might never want to see it and they should have the ability to customize their interface accordingly.

8 LIMITATIONS

An important factor to keep in mind is that we conducted a lab study, as opposed to an online study with real players as the one of the three-star system of FoldIt [17]. Although there are advantages to conducting lab studies (e.g. more control over the experiment duration and set up, possibility of answering questions, ability to observe and interview participants, etc.), there are definitely some drawbacks. One major challenge is to recruit participants who can faithfully represent the usual human computing game players. In the context of a real human computing game for example, one could expect that a portion of the participants would be more interested in helping to solve a problem (scientific or not) and have limited video game experience, whereas others would be more interested in the game itself or the social aspect of it (competing or collaborating with others). Even with a careful design of the recruitment posters, it is very hard to ensure a good representative sample, and that can obviously affect some of the results.

Another limitation of this study is that we have used a mock-up leaderboard with 9 simulated player scores. This leaderboard does not represent a real online game context, in which the top players might eventually be unreachable. The objective of the leaderboard as a gamification technique is to stimulate participants to perform better in a more competitive environment [18, 21, 24], which is hard to replicate in a controlled lab experiment. Also, since each experiment had a duration of roughly 45 minutes, it was difficult to reproduce and analyze the benefits that a leaderboard could have on a long term experiment. Different techniques such as connecting the leaderboard with social media so that the players can play with their friends could be more effective too.

Finally, we did not test different locations for the game features on the screen. Locations of the game features help the players to focus on a particular point which could make a difference on their performance. A player mentioned for example that the leaderboard was distracting, so maybe it was taking too much space on the screen. A future study could investigate the proper sizes and locations for showing the leaderboard, the timer and the top score.

9 CONCLUSION

The goal of this study was to investigate the effect of different game mechanics (timer, top score and live leaderboard) on player performance and motivation in a human computing game. We observed that the timer and the top score had significant effects on players' scores and completion times. Presenting an achievable top score resulted in more optimal solutions (*i.e.* solutions with fewer operations), but at the expense of the completion time, whereas the timer had the opposite effect on players. The live leaderboard, along with a context-dependent points system, wasn't significantly affecting the scores and completion times, but we did observe an almost significant interaction effect with the timer on the score (the scores being worse when the timer and leaderboard were present). Although the top score seemed to be the most appreciated and motivating game element, our interviews with the players did not highlight a clear consensus on the three game mechanics. These findings tend to suggest that customizability of the game interface (e.g. showing or not a live leaderboard) and/or letting players choose

between different game modes (with or without certain mechanics) might be preferable to promote player motivation in a human computing game. Exploring the generalizability of these findings on a larger sample size of regular human computing game users by conducting an online study would be an interesting direction for future work.

Another important question is what motivates users to participate in and stay engaged with human computing, crowdsourcing and citizen science applications in general (see [13] for an example of such a study in the context of the FoldIt game). Human computing, crowdsourcing and citizen science platforms should not only be focused on companies or scientists collecting data. They must also be concerned about citizens actively participating in problem solving/research, collaborating with other participants and/or researchers, and learning something valuable from the experience. Studying how to properly give back to the users, how to improve learning outcomes from participating in these platforms, how to promote collaborative work, etc. are all crucial questions that need to be answered in future work.

REFERENCES

[1] Graham Alvare and Richard Gordon. 2015. CT brush and CancerZap!: two video games for computed tomography dose minimization. *Theoretical Biology and Medical Modelling* 12, 1 (2015), 7.

[2] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. 2017. GenBank. *Nucleic acids research* 45, D1 (2017), D37–D42.

[3] François Bouchy, Maxime Marmier, and Oliver Turner. 2018. Light-Curve Classifications Scrutinized by UNIGE Astronomers. http://mmos.ch/news/2018/01/26/light-curve-classifications-scrutinized-by-unige-astronomers.html

[4] Jason T Bowey, Max V Birk, and Regan L Mandryk. 2015. Manipulating leaderboards to induce player experience. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. ACM, 115–120.

[5] Anne Bowser, Derek Hansen, Yurong He, Carol Boston, Matthew Reid, Logan Gunnell, and Jennifer Preece. 2013. Using gamification to inspire new citizen science volunteers. In *Proceedings of the first international conference on gameful design, research, and applications*. ACM, 18–25.

[6] Anne E Bowser, Derek L Hansen, Jocelyn Raphael, Matthew Reid, Ryan J Gamett, Yurong R He, Dana Rotman, and Jenny J Preece. 2013. Prototyping in PLACE: a scalable approach to developing location-based apps and games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1519–1528.

[7] CCP. 2017. EVE Online: Project Discovery. https://www.eveonline.com/discovery/

[8] Irene Celino, Dario Cerizza, Simone Contessa, Marta Corubolo, Daniele DellAglio, Emanuele Della Valle, and Stefano Fumeo. 2012. Urbanopoly–A Social and Location-Based Game with a Purpose to Crowdsource Your Urban Data. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 910–913.

[9] Xin Chen, Jie Zheng, Zheng Fu, Peng Nan, Yang Zhong, Stefano Lonardi, and Tao Jiang. 2005. Assignment of orthologous genes via genome rearrangement. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 2, 4 (2005), 302–315.

[10] David Codish and Gilad Ravid. 2014. Personality based gamification-Educational gamification for extroverts and introverts. In *Proceedings of the 9th CHAIS Conference for the Study of Innovation and Learning Technologies: Learning in the Technological Era*, Vol. 1. 36–44.

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

[12] Seth Cooper, Adrien Treuille, Janos Barbero, Andrew Leaver-Fay, Kathleen Tuite, Firas Khatib, Alex Cho Snyder, Michael Beenen, David Salesin, David Baker, et al. 2010. The challenge of designing scientific discovery games. In *Proceedings of the Fifth international Conference on the Foundations of Digital Games*. ACM, 40–47.

[13] Vickie Curtis. 2015. Motivation to participate in an online citizen science game: A study of Foldit. *Science Communication* 37, 6 (2015), 723–746.

[14] 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*. ACM, 97–101.

[15] Chris Drogaris, Akash Singh, Elena Nazarova, Lance Zhou, Alex Kawrykow, Gary Roumanis, Alfred Kam, Mathieu Blanchette, and Jérôme Waldispühl. 2018. Phylo. https://phylo.cs.mcgill.ca/

[16] Alexandra Eveleigh, Charlene Jennett, Stuart Lynn, and Anna L Cox. 2013. "I want to be a Captain! I want to be a Captain!": Gamification in the Old Weather Citizen Science Project. In *Proceedings of the first international conference on gameful design, research, and applications*. ACM, 79–82.

[17] Jacqueline Gaston and Seth Cooper. 2017. To three or not to three: Improving human computation game onboarding with a three-star system. In *Proceedings of the 2017 CHI conference on Human Factors in Computing Systems*. ACM, 5034–5039.

[18] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences (HICSS)*. IEEE, 3025–3034.

[19] Simone Hantke, Florian Eyben, Tobias Appel, and Björn Schuller. 2015. iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 891–897.

[20] Amazon Mechanical Turk Inc. 2018. Amazon Mechanical Turk. https://www.mturk.com/

[21] Yuan Jia, Yikun Liu, Xing Yu, and Stephen Voida. 2017. Designing Leaderboards for Gamification: Perceived Differences Based on User Ranking, Application Domain, and Personality Traits. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1949–1960.

[22] Alexander Kawrykow, Gary Roumanis, Alfred Kam, Daniel Kwak, Clarence Leung, Chu Wu, Eleyine Zarour, Luis Sarmenta, Mathieu Blanchette, Jérôme Waldispühl, et al. 2012. Phylo: a citizen science approach for improving multiple sequence alignment. *PloS one* 7, 3 (2012), e31362.

[23] Jinseop S Kim, Matthew J Greene, Aleksandar Zlateski, Kisuk Lee, Mark Richardson, Srinivas C Turaga, Michael Purcaro, Matthew Balkam, Amy Robinson, Bardia F Behabadi, et al. 2014. Space–time wiring specificity supports direction selectivity in the retina. *Nature* 509, 7500 (2014), 331.

[24] Richard N Landers, Kristina N Bauer, and Rachel C Callan. 2017. Gamification of task performance with leaderboards: A goal setting experiment. *Computers in Human Behavior* 71 (2017), 508–515.

[25] Edith Law and Luis Von Ahn. 2009. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1197–1206.

[26] Jeehyung Lee, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Snehal Gaikwad, Sungroh Yoon, Adrien Treuille, et al. 2014. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences* 111, 6 (2014), 2122–2127.

[27] 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*. ACM, 89–98.

[28] Amaia Salvador, Axel Carlier, Xavier Giro-i Nieto, Oge Marques, and Vincent Charvillat. 2013. Crowdsourced object segmentation with a game. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*. ACM, 15–20.

[29] Jürgen Scheible, Ville H Tuulos, and Timo Ojala. 2007. Story Mashup: design and evaluation of novel interactive storytelling game for mobile and web users. In *Proceedings of the 6th international conference on Mobile and ubiquitous multimedia*. ACM, 139–148.

[30] Jesse Schell. 2019. *The Art of Game Design: A book of lenses*. AK Peters/CRC Press.

[31] Akash Singh, Faizy Ahsan, Mathieu Blanchette, and Jérôme Waldispühl. 2017. Lessons from an Online Massive Genomics Computer Game.. In *HCOMP*. 177–186.

[32] Akash Singh, Chris Drogaris, Elena Nazarova, Mathieu Blanchette, Jérôme Waldispühl, Anjum Ibna Matin, and O Tremblay-Savard. 2017. A human-computation platform for multi-scale genome analysis. *Association for the Advancement of Artificial Intelligence (www. aaai. org). Retrieved from https://www. humancomputation. com/2017/papers/100_human-computation-platform. pdf* (2017).

[33] Kristin Siu, Matthew Guzdial, and Mark O Riedl. 2017. Evaluating singleplayer and multiplayer in human computation games. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*. ACM, 34.

[34] Kristin Siu and Mark O Riedl. 2016. Reward systems in human computation games. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. ACM, 266–275.

[35] Kristin Siu, Alexander Zook, and Mark O Riedl. 2014. Collaboration versus competition: Design and evaluation of mechanics for games with a purpose. In *FDG*.

[36] Kristin Siu, Alexander Zook, and Mark O Riedl. 2017. A framework for exploring and evaluating mechanics in human computation games. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*. ACM, 38.

[37] Jens Jakob WH Sørensen, Mads Kock Pedersen, Michael Munch, Pinja Haikka, Jesper Halkjær Jensen, Tilo Planke, Morten Ginnerup Andreasen, Miroslav Gajdacz, Klaus Mølmer, Andreas Lieberoth, et al. 2016. Exploring the quantum speed limit with computer games. *Nature* 532, 7598 (2016), 210.

[38] Devin P Sullivan, Casper F Winsnes, Lovisa Åkesson, Martin Hjelmare, Mikaela Wiking, Rutger Schutten, Linzi Campbell, Hjalti Leifsson, Scott Rhodes, Andie Nordgren, et al. 2018. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature biotechnology* (2018).

[39] Olivier Tremblay-Savard, Alexander Butyaev, and Jérôme Waldispühl. 2016. Collaborative Solving in a Human Computing Game Using a Market, Skills and Challenges. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. ACM, 130–141.

[40] Ville H Tuulos, Jürgen Scheible, and Heli Nyholm. 2007. Combining web, mobile phones and public displays in large-scale: Manhattan story mashup. In *International Conference on Pervasive Computing*. Springer, 37–54.

[41] Marc Van Kreveld, Maarten Löffler, and Paul Mutser. 2015. Automated puzzle difficulty estimation. In *2015 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 415–422.

[42] 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, 319–326.

[43] Luis Von Ahn, Ruoran Liu, and Manuel Blum. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 55–64.

[44] Greg Walsh and Jennifer Golbeck. 2010. Curator: a game with a purpose for collection recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2079–2082.

[45] Inc. Weblabcenter. 2018. Microworkers. https://microworkers.com/