

Interplay of Game Incentives, Player Profiles and Task Difficulty in Games with a Purpose

Gloria Re Calegari and Irene Celino

Cefriel – Politecnico of Milano, Viale Sarca 226, 20126 Milano, Italy
 {gloria.re, irene.celino}@cefriel.com

Abstract. How to take multiple factors into account when evaluating a Game with a Purpose? How is player behaviour or participation influenced by different incentives? How does player engagement impact their accuracy in solving tasks? In this paper, we present a detailed investigation of multiple factors affecting the evaluation of a GWAP and we show how they impact on the achieved results. We inform our study with the experimental assessment of a GWAP designed to solve a multinomial classification task.

1 Introduction

Games with a Purpose [1] are a well-known Human Computation approach [2] to encourage users to execute tasks with an entertaining reward. While several metrics are proposed in literature to evaluate the ability of GWAPs to achieve their intended purpose, there is a large number of factors that influences their success and effectiveness.

In order to fully understand the strengths as well as the weaknesses of a GWAP, we propose an approach that takes into account *player characteristics* (reliability, participation, behaviour and accuracy), *game aspects* (playing incentive, playing style and game nature) and *features of the task* to be solved (level of difficulty and variety). Our goal is to investigate the interplay between those different factors, by proposing a multi-faceted analysis framework that allows for a deep assessment and understanding of the efficacy of a GWAP to achieve its purpose. We apply the proposed framework to a specific GWAP to show the empirical results and the insights that can be drawn through our approach.

The original contributions of this paper are: (1) an extension of traditional GWAP metrics to take temporal evolution and incentive effects into account; (2) a comparison of engagement metrics and engagement profiles with non-gaming citizen science; and (3) the definition of GWAP-specific engagement profiles and their interplay with different factors (incentive, task difficulty and task variety).

The remainder of the paper is organized as follows: Section 2 illustrates the main related work; Section 3 gives details about the GWAP that we use to exemplify our approach; in the following sections, we propose different evaluation methods, by extending state-of-the-art metrics: global GWAP metrics and interplay with incentive are adopted in Section 4, Section 5 offers a comparison with citizen science user engagement profiles and Section 6 proposes new GWAP

player profiling driven by measures of participation and accuracy; finally, Section 7 concludes the paper.

2 Related work

The basic metrics to evaluate GWAPs [1,2,3] are global indicators computed as means over the entire data; while effective in summarizing the behaviour of GWAP players, those are very simple measures that do not tell the entire story: an analysis of data distribution and temporal evolution is usually required to get a deeper understanding of a GWAP.

Some work exists on cross-feature analysis of GWAPs [4] and similarly on citizen science [5] and crowdsourcing [6]; our goal is to contribute to making such evaluation easier to replicate and reproduce.

Participation incentives are usually classified as intrinsic or extrinsic motivation [7]. Some comparative analysis of incentives exists for GWAPs [8], especially in contrast to different methods like micro-working [9,10,11] or machine learning [12]. The effect of competition and tangible rewards on participation and quality of results has also been explored, both in the context of GWAPs [13] and online citizen science campaigns [14], revealing the pros and cons of designing different motivation mechanisms.

Other metrics to evaluate GWAPs can be borrowed from studies of social community [15] and citizen science evolution [16]; in those cases, however, user participation's "success" is measured through simple indicators like number of participants and contributions, while a deeper investigation is needed to assess the effectiveness of participation. Behavioural studies in HCI research have investigated volunteer characterization in citizen science, defining engagement metrics and profiles [17,18], which may or may not apply to GWAP players.

In the context of (paid) crowdsourcing, assessment is usually conducted in relation to micro-work platforms [19], in which important features are related to cost minimization [20,21] which is out of scope with respect to our work.

While Games with a Purpose are a well-known and widely adopted human computation method to involve users in task solution, a comprehensive assessment of their ability to address their "purpose" needs to take into account multiple factors affecting the game and the players. We therefore propose a multi-faceted analysis framework for GWAPs that includes game aspects, player characteristics and task features, with specific focus on the effect of game incentives on the overall GWAP efficacy.

3 Use Case: the Night Knights GWAP

The GWAP that we will use as running example is Night Knights, an online game for the multinomial classification of images¹. Pictures come from a massive public-domain dataset provided by NASA and they can be classified according to

¹ Cf. <https://www.nightknights.eu/>.



Fig. 1. Night Knights: the gameplay

six different categories depending on their visual content. The classified images – in particular those labeled with three of the six categories – are then used in a subsequent scientific workflow in the field of astronomy and environmental sciences to measure light pollution effects (cf. [12]).

The GWAP is inspired by the ESP game [3], because users play in random pairs according to an output-agreement mechanism [1]. The game adopts a repeated labeling approach [22] by asking different players to classify the same image; conversely, the same image is never given twice to the same player. Night Knights is built on top of our open source software framework for GWAPs [23].

The players visualize a picture and six buttons reporting the six possible categories (cf. Figure 1); the labeling task is therefore executed by clicking on the category that better fits the picture content. Each game round lasts one minute, during which players can classify as many images as they can (as detailed in the following, on average 15 pictures are played per round); each time the two players agree, they gain points and level up in the game leaderboard; some badges are also assigned in special conditions as additional game intrinsic incentives.

Players’ contributions are aggregated through an incremental truth inference algorithm [24] that (1) processes inputs as soon as they are provided, (2) weights players’ answer with a round-specific reliability measure [25] taking into account players’ answers on control tasks (for which the “true” solution is known), and (3) dynamically adjusts the number of required contributions. Our truth inference approach accounts for the very nature of GWAPs, in which usually there is no “deadline” for contributing, players’ varying attention can impact answer

quality and task difficulty needs a dynamic estimation of the required number of repeated labeling.

In this paper, we use the data collected through Night Knights. The game was released in February 2017 and then it was more extensively advertised for a related competition whose winner joined the 2017 Summer Expedition to observe the Solar Eclipse in USA. The competition lasted about one month, from mid June to mid July 2017, and was addressed to all EU University students. After the end of the competition, the game has still been available online, but without any additional advertising. Overall, the data we analyse was collected in 9 months, one month of competition and 4 months before and after it².

In the following experimental sections, we apply a set of assessment methods on this game data. On the one hand, we exemplify the analyses we propose for a thorough multi-faceted assessment of GWAPs; on the other hand, we provide concrete results from the evaluation of Night Knights, which are – at least partially – typical of GWAPs.

4 Extending GWAP metrics

The main metrics adopted in literature [2] to evaluate GWAPs are: **throughput**, computed as the average number of solved task per unit of time, **average life play** or ALP, i.e. the average time spent by each user playing the game, and **expected contribution** or EC, measured as average number of tasks solved by each player. A task is solved when player contributions, aggregated by the truth inference algorithm [26], output a “true” solution. Those indicators are global measures, as they are computed as mean values over the entire GWAP use. Hereafter, we extend this analysis by assessing the *influence of different game incentives* and the *evolution over time* of game-play and engagement.

In particular, we investigate how player participation and GWAP results change with and without an extrinsic motivation such as a tangible reward [7]. We analyse incentive effect in terms of both general statistics and specific metrics adopted in GWAP evaluation. We show that users participation can be highly influenced by the presence of an extrinsic motivation.

4.1 [Q1] How do user participation and GWAP results change with different incentives?

In 9 months, Night Knights managed to engage about 650 users that played a substantial amount of time and classified almost 28,000 photos (cf. Table 1).

Measuring the main metrics in the three periods (*before*, *during* and *after* the competition), we notice a significant increase of player participation during the competition, both in terms of given contributions and classified images (one order of magnitude higher with the additional incentive in both cases). This difference is clearly highlighted in Figure 2, which shows the temporal evolution of the number of images classified per day. The difference between throughput, ALP

² Data is available with a CC-BY license at <http://ckan.stars4all.eu/>.

	Before	During	After
time span (months)	4	1	4
classified images	1,830	24,600	1,300
contributions	13,000	187,600	3,600
users	285	174	174
total play time (hours)	65	471	29
throughput (tasks/hour)	69	212	113
ALP (mins/user)	5.5	65	4
EC (tasks/user)	6.4	141	7.5

Table 1. Experimental results in the three periods (before, during and after the introduction of the extrinsic motivation)

and EC in the competition and non-competition periods is statistically significant (t-test or Wilcoxon rank sum test at the 0.01 significance level). Also the play time significantly increases *during* the competition period, as demonstrated by the ALP metrics which reaches values over 65 minute/player (cf. Table 1).

Those results prove that providing a tangible reward to players can make them contribute more efficiently, speeding up the classification process (higher throughput), engaging them for a longer time (higher ALP), and ensuring a larger contribution rate to the human computation task (higher EC). As a global result, more tasks get solved.

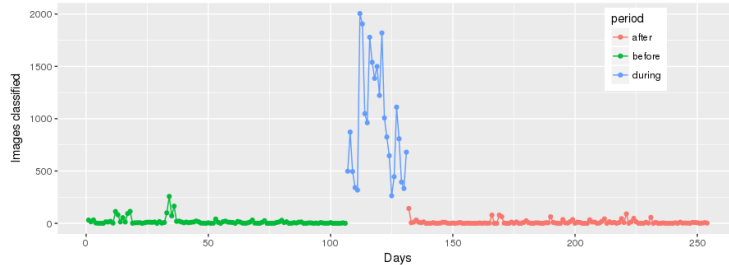


Fig. 2. Number of images classified per day in the three periods

4.2 [Q2] Do the extrinsic reward effects last over time?

Adding a tangible prize to a game does not seem to ensure lasting effects. In Night Knights, looking deeper in the *before* and *after* periods in Table 1, we do not notice substantial differences in terms of classification and participation rate. The metrics of the *before* period are slightly higher, probably due to the fact that more users tried the game, attracted by advertising campaigns (small peaks in Figure 2) and by the novelty of the game.

Given this similarity, in our analysis we think it worth distinguishing only between *intrinsic motivation* periods (e.g., Night Knights *before* and *after* periods together, when users play only to have fun) and *extrinsic motivation* periods (e.g., the *during* phase of Night Knights, with the tangible and valuable reward).

4.3 [Q3] Does playing style change with the incentive?

Defining *contribution speed* the number of images played in each round, we check if also this metrics is influenced by a tangible reward.

As explained in Section 3, each round in Night Knights lasts one minute and each user is asked to classify one image at a time, so users have to be quick and classify as many images as possible to increase their score and being successful in the game. Given the image loading time, connection delays and waiting time for the other player’s answer, we estimate that in this case classifying each image takes at least 3–5 seconds, which means 12–20 photos per round.

As Figure 3 shows, in the *extrinsic motivation* period, the contribution speed follows a normal distribution centered around 15 photos/round, while, in the *intrinsic motivation* phase, the distribution is flat and most players played less than 10 images/round. This indicates that, during the competition, all players did their best to classify as many images as possible, reaching a median value of 15 that coincides with the estimated image classification time. On the other hand, in the *intrinsic motivation* period, people play the game in a more “relaxed” way, just to try and explore it, taking more time to answer.

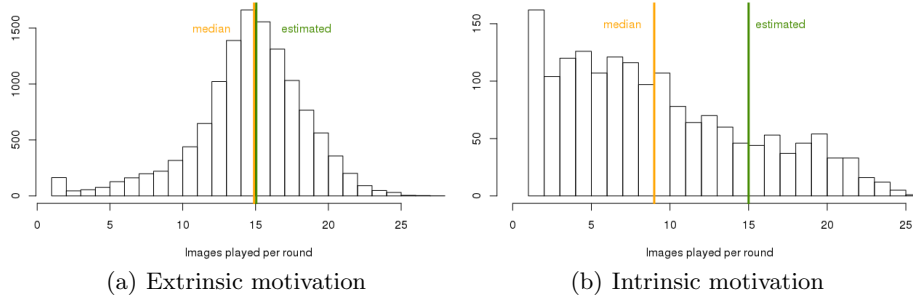


Fig. 3. Distribution of the number of images played in each round

5 Applying Citizen Science Engagement profiles

As a first step to the assessment of player behaviour, we adopt the *engagement metrics* proposed by [17]: **activity ratio**, number of days a user plays a game divided by the total number of days the user remains linked to the game; **daily devoted time**, average time (e.g. in hours) a user plays the game in each active day; **relative active duration**, ratio of days during which a player remains linked to the game and the total number of days since the player joined the game until the day the game is over (this metric can be computed only if a “game end” is envisaged, which is not always the case in GWAPs); and **variation in periodicity**, standard deviation of the intervals between each pair of non-consecutive active days. Computing those metrics for each player and then

applying clustering techniques leads to the identification of *engagement profiles*. Our goal is to assess if the profiles recognized in citizen science literature with respect to volunteer behaviour are also detected in GWAP player behaviour and if player profiles are affected by game incentives. Indeed, we expect player behaviour to differ from volunteer engagement.

5.1 [Q4] How does GWAP behaviour compare to traditional citizen science engagement?

The mean values (and in brackets standard deviation) of the four main *engagement metrics* defined by [17] are shown in Table 2. For Night Knights, we distinguish the global values and those measured during the competition only (extrinsic motivation period); for comparison, we also report the values for the citizen science initiatives illustrated in [17,18]. Daily devoted time for Night Knights is measured by approximation, multiplying the number of game rounds per 1-minute duration (the actual time is higher, because players also browse leaderboards, badges, played pictures, etc.); relative active duration is computed only during the competition time, where a “project finish time” is defined with the contest deadline.

We observe that Night Knights players display quite a different behaviour with respect to volunteers: they show a 2-3 times higher activity ratio, and also consistently higher values for daily devoted time and relative active duration; this may mean that GWAP players tend to contribute in a more regular manner than volunteers. Focusing on the competition, those metrics also show a clear increase in engagement, with a significantly lower value of variation in periodicity, which suggests that the limited-time contest period stimulates players to access the game even more frequently and regularly.

Clustering players to identify engagement profiles does not give the same results as in the cited citizen science analyses [17,18]. Cross-validation between different methods (within groups sum of squares and Silhouette statistics) suggests an optimal clustering with 3 groups. Applying both agglomerative hierarchical clustering and K-means clustering yields to similar and very unbalanced grouping, with one big cluster (around 90% of players) roughly corresponding

	Night Knights		MW	GZ	WI
	global	compet.	[17]	[17]	[18]
Activity ratio	0.96 (0.17)	0.95 (0.16)	0.40 (0.40)	0.33 (0.38)	0.32 (0.35)
Daily devoted time	0.68 (1.94)	1.80 (3.30)	0.44 (0.54)	0.32 (0.40)	–
Rel. active duration	–	0.54 (0.35)	0.20 (0.30)	0.23 (0.29)	0.43 (0.44)
Var. in periodicity	14.53 (17.9)	2.53 (2.12)	18.27 (43.3)	25.23 (49.2)	5.11 (5.36)

Table 2. Engagement metrics (mean values and standard deviation in brackets): comparison of Night Knights (global values and competition-only metrics) with citizen science campaigns (MW: Milky Way, GZ: Galaxy Zoo, WI: Weather-it).

to the *hardworker* profile (high activity ratio and low variation in periodicity); the remaining players are grouped in a small cluster that we can name “*focused*” *hardworkers* (similar to hardworkers but with higher daily devoted time) and another small cluster that does not clearly correspond to known profiles (low values of all metrics, but higher variation in periodicity). The spasmodic, persistent, lasting and moderate profiles defined in [17] are not observed. This can be interpreted as another difference between players and volunteers engagement, with game users either heavily playing and contributing, or simply trying out the game without being actually engaged.

5.2 [Q5] What does player behaviour tell about the game nature?

If we also evaluate user engagement in terms of when players participated, i.e. for how long they played the game, from the first to the last played round, we discover that only few users played the game both in the intrinsic motivation and extrinsic motivation periods; in particular, only 13 users played both *before* and *during* the competition and only 17 users became aware of the existence of the game during the competition and went on playing it *after* its end.

In addition, by analysing the users’ total active time (difference between the last and the first time a user played the game), we discover that most of the users played for a very short amount of time; 75% of players used the game for less than 5 minutes and only the 10% played for more than a day.

These statistics are not surprising, because they are strong indicators of the game nature, which is a so-called *casual game*. Casual games are usually designed to be played in short bursts of a few minutes and then set aside. By their very nature, casual games target the short free/leisure time between the myriad of everyday tasks, such as between work and domestic obligations or between attention and distraction [27]. Regarding the overall time spent playing mobile games, the literature shows that an average gamer spends every day approximately 24 minutes playing games on mobile devices, with *heavy gamers* spending about 1 hour/day and *light gamers* about 2 minutes/day [28].

6 Defining GWAP Engagement profiles

Given that volunteer profiles in citizen science do not seem to suitably describe GWAP players, we focus our investigation on two additional main metrics, player accuracy and player participation, more closely related to human computation, and analyse their interplay with different factors, like game incentive, task difficulty and task variety. The goal is to uncover GWAP-specific user behaviours and to identify *GWAP-specific player profiles*.

Player accuracy is measured ex-post by counting how many tasks each user correctly solved over the total number of tasks he/she played with; in this context, “correct” refers to the final task solution computed by the truth inference algorithm. Accuracy takes values between 0 and 1 and corresponds to the worker precision or labeling quality metrics used in crowdsourcing literature

(e.g. [26]). **Player participation** is measured as the total number of contributions given by each user in the game rounds he/she played. While there are of course alternative ways to measure participation (e.g., number of game rounds, total played time), we prefer to consider the number of contributions, since this indicator is more closely related to the “task” execution and the game purpose.

6.1 [Q6] What kind of GWAP player profiles can be identified?

Referring again to Night Knights data, we plot each user as a data point along participation and accuracy axes (cf. Figure 4). To divide players into groups, we applied clustering as in Section 5, but – at least in the case of Night Knights – the results put 98-99% of players in the same cluster, placing only “outliers” in the other clusters. Therefore, to define GWAP-specific profiles, we propose to simply set separation thresholds on the two axes dividing the space into quadrants; more specifically, we adopt the *median* as separation value, which is a commonly used measure and robust statistic. While this definition is arbitrary, it is also data-independent, thus the proposed approach can be adopted to analyse and compare different GWAPs without loss of generality.

The thresholds calculated on the Night Knights dataset are 12 contributions for the x-axis and 0.87 accuracy for the y-axis. The median value for participation roughly corresponds to the separation between those who played just a couple of game rounds from those who were more deeply engaged (cf. Section 4). The median accuracy value is quite high and this is a good sign about the GWAP efficacy to achieve its purpose; in other cases, when a specific minimum value of accuracy is required, the threshold choice could be driven by domain-specific consideration instead of being identified by the median.

By using this approach, the investigation space is divided into areas that represent different “behavioral” profiles as follows. Along the accuracy axis, we obtain two profiles: *accurate players*, i.e. players with an accuracy higher than the median, and the remaining *inaccurate players* (cf. Figure 5-a). Along the

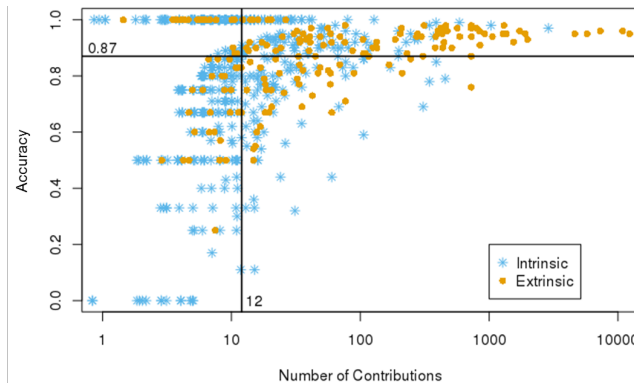


Fig. 4. Players’ participation vs. accuracy and median values

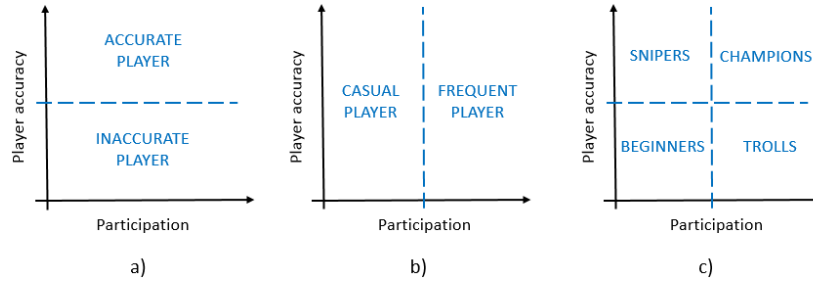


Fig. 5. Definition of GWAP-specific player profiles

participation axis (cf. Figure 5-b), we define *casual players* those who contribute less than the median, and *frequent players* the most addicted and loyal contributors. Considering both dimensions, we define four profiles (cf. Figure 5-c):

- *Beginners* (bottom-left): this is the set of users that play the game for a short period of time, just for curiosity; this kind of players gives only few contributions with low accuracy.
- *Snipers* (top-left): users that are very accurate in their contributions but they contribute only a little. Ideally, they should be motivated to become champions, since their contributions are valuable.
- *Champions* (top-right): this is the most desirable category of players, since they have high level of participation with very high accuracy.
- *Trolls* (bottom-right): this is the category of less desirable users, since they give a lot of inaccurate contributions; having a lot of *Trolls* in the game either makes the classification process longer, since it is harder to reach an agreement, or even leads to undesired results.

Observing again Night Knights data, we can also quantitatively analyse the effect of game incentive on the profile composition (cf. Figure 6). With extrinsic motivation, most users (53%) acted as champions, and this share is much higher than in the total (32%). On the other hand, with the intrinsic motivation only,

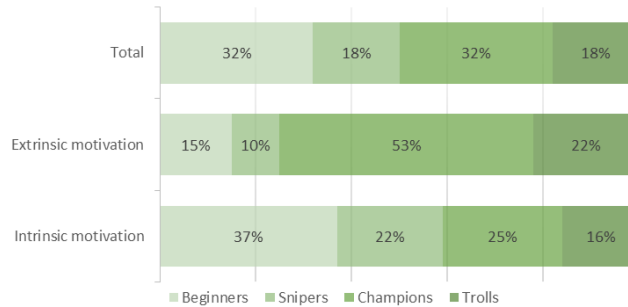


Fig. 6. Distribution of players between profiles, in total and with different incentives

the presence of champions was lower, only 25%. This difference may indicate that the different incentives lead to different user behaviour; the presence of tangible rewards can engage users for a longer time and can motivates them to contribute with more effort and attention.

With intrinsic motivation, also the percentages of snipers was higher than the average. The largest group of users in the intrinsic motivation period, however, was beginners (37%): probably this happened because they tried the game just for curiosity or to understand how the game works, without paying too much attention to the answers they gave. As expected, the number of beginners was very low with the extrinsic motivation, since they had a clear goal to play the game. Fortunately, the percentages of trolls were low in both periods. This means that the Night Knights game succeeded in avoiding too many spammers that could have made the classification process longer or more inaccurate.

While the above results are specific to Night Knights, the profile analysis can be applied to any other GWAP; indeed, examining the composition of a GWAP player population can reveal different behaviour and inform game re-design.

Finally, we would like to point out an insight that is not immediately evident in Figure 4: since the players on the right part of the plot are those who contributed more, if we sum the contributions from the four profiles, we obtain the figures in Table 3. In the case of our GWAP, therefore, the large majority of contributions comes from the most active and accurate players, which is reassuring with respect to the achievement of the game purpose.

	Beginners	Snipers	Champions	Trolls
Task contributions	0.7%	0.4%	95.9%	3.0%

Table 3. Distribution of contributions across players profiles

In the following, we analyse the interplay between player accuracy and player participation by taking into account additional factors. More specifically, we check if there is a statistically significant difference between the mean accuracy of *casual* and *frequent players* with respect to some control variables, namely the incentive type, the task difficulty and the task variety.

6.2 [Q7] Does player behaviour change with different incentives?

To answer this question, we check for mean difference in accuracy for casual and frequent players in the intrinsic and extrinsic motivation periods.

In Night Knights, the average accuracy of the *frequent players* is higher than the one of *casual players* in both periods, as shown in the first two boxplots of Figure 7; this difference is also significant from a statistically point of view (p-value of the t-test less than 0.05). We also notice a mean accuracy increase of about 10% when a tangible rewards is present (from 0.74 to 0.81 for casual and from 0.83 to 0.90 for frequent): since during the competition users were

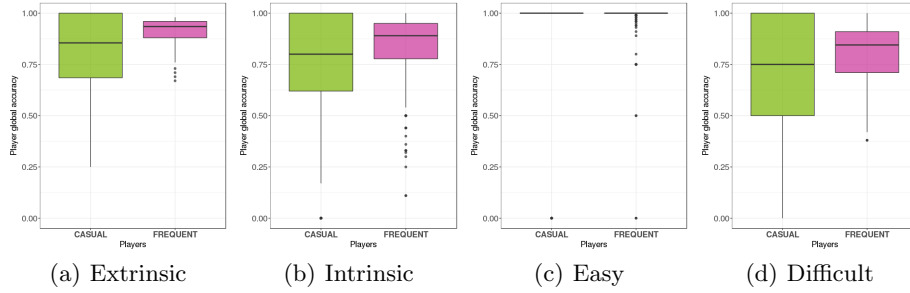


Fig. 7. Accuracy distribution of *casual* and *frequent* players with different incentives (a and b) and with different task difficulty (c and d). The difference between players’ profiles is statistically significant in all cases except for easy tasks.

encouraged to play to win the prize, they paid more attention to the image classification, raising also the answers’ quality.

This may indicate that in GWAPs *frequent players* contribute in a more accurate way than *casual* ones, and that extrinsic motivation has a positive impact on accuracy.

6.3 [Q8] Does player behaviour change with task difficulty?

We define *task difficulty* as the number of different users needed to solve it (the higher the number, the harder the task); this is because our incremental truth inference algorithm (cf. Section 3) dynamically estimates the number of contributions required to solve a task. We split the images in two sets based on their difficulty and we check if this impacts player behaviour.

For Night Knights, we marked as “easy” the images that requires only 4 contributions (the minimum number to reach an agreement according to our domain experts), and as “difficult” those that required more contributions. “Easy” images are 58% of all classified images, while the number of contributions required to classify “difficult” images ranges from 5 to 17.

As shown in the (c) and (d) boxplots in Figure 7, accuracy on “easy” images is almost the same between *casual* and *frequent players* (indeed, the difference in mean accuracies is not statistically significant). On the contrary, this difference is statistically significant for “difficult” images (mean accuracy is 0.84 for *frequent players* and 0.68 for *casual players*).

Those results suggest a *learning effect* in GWAPs: the more a user plays the game, the more he/she understands the task to be solved, thus increasing his/her accuracy and consequently also result quality.

6.4 [Q9] Does player behaviour change with task variety?

Since Night Knights aims to solve a multinomial classification task, we investigate whether there is any evident phenomenon related to the different image

	Black City Stars Aurora ISS None					
Casual	0.69	0.88	0.57	0.74	0.63	0.70
Frequent	0.79	0.91	0.68	0.77	0.77	0.77

Table 4. Mean accuracy of *casual* and *frequent* players with images of different categories. The difference is not statistically significant for any of the categories.

categories. Therefore, we compute again the accuracies of the two groups of casual and frequent players in classifying the 6 output classes. We summarize the mean accuracy values in Table 4.

Applying the t-test to check if the mean accuracy is different for the two players’ profiles, we cannot reject the null hypothesis. This may mean that any player is equally able/unable to distinguish the different categories, independently of his/her level of participation; indeed, in our GWAP, there is no need for background- or domain-specific knowledge to play the game. This analysis can help in identifying the need for training or expert knowledge of GWAP players.

On the other hand, the mean accuracy values change a lot across different categories, spanning between 0.57 and 0.91. This is also explained by the different distribution of easy/difficult tasks across the variety of classes, as shown in Figure 8. Indeed, some categories are intrinsically more difficult to classify than others, but Table 4 shows that this complexity related to task variety is equally perceived by players with low and high levels of participation.

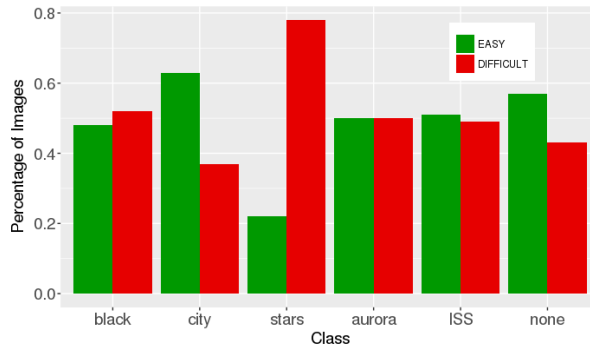


Fig. 8. Distribution of easy/difficult tasks across different image categories.

7 Conclusions

In this paper, we presented an investigation of the interplay of different factors in the evaluation of GWAP results. More specifically, we focused on the profiling of players according to different user metrics and we studied the influence of game incentive and task characteristics.

To inform our discussion, we described the results of such multi-dimensional analysis over the data collected by a GWAP for multinomial classification of images. While some of our considerations result from the quantitative analysis of a single game, and are not *per se* generalizable, we believe that the proposed approach is replicable to evaluate any other GWAP. We believe that such deeper analysis is an important (and sometimes neglected) investigation to understand players' behaviour, to evaluate the impact of various factors on reliability and quality, and finally to assess the ability of GWAPs to achieve their intended purpose and its sustainability over time.

Finally, we would like to point out that, even when player participation is limited in time, a classification GWAP can be used to build a reasonably large training set to be used in traditional machine learning settings to train classifiers for larger-scale labeling. In our previous work, we showed that humans and machines indeed agree on image classification for the Night Knights dataset [12].

Acknowledgments

This work is partially supported by the STARS4ALL project (H2020-688135), co-funded by the European Commission. We thank all the Night Knights players who contributed to the classification task solution and allowed us to perform this work.

References

1. Von Ahn, L., Dabbish, L.: Designing games with a purpose. *Communications of the ACM* **51**(8) (2008) 58–67
2. Law, E., Ahn, L.v.: Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **5**(3) (2011) 1–121
3. Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2004) 319–326
4. Singh, A., Ahsan, F., Blanchette, M., Waldisp"uhl, J.: Lessons from an online massive genomics computer game. In: *Proceedings of the Fifth Conference on Human Computation and Crowdsourcing (HCOMP 2017)*. (2017)
5. Sauermann, H., Franzoni, C.: Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences* **112**(3) (2015) 679–684
6. Yang, J., Redi, J., Demartini, G., Bozzon, A.: Modeling task complexity in crowdsourcing. In: *Fourth AAAI Conference on Human Computation and Crowdsourcing*. (2016)
7. Ryan, R.M., Deci, E.L.: Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology* **25**(1) (2000) 54–67
8. Prestopnik, N., Crowston, K., Wang, J.: Gamers, citizen scientists, and data: Exploring participant contributions in two games with a purpose. *Computers in Human Behavior* **68** (2017) 254–268
9. Thaler, S., Simperl, E., Wolger, S.: An experiment in comparing human-computation techniques. *IEEE Internet Computing* **16** (2012) 52–58
10. Feyisetan, O., Simperl, E., Van Kleek, M., Shadbolt, N.: Improving paid micro-tasks through gamification and adaptive furtherance incentives. In: *Proceedings of the 24th International Conference on World Wide Web. WWW '15, Republic and Canton of Geneva, Switzerland, International World Wide Web Conferences Steering Committee* (2015) 333–343

11. Feyisetan, O., Simperl, E.: Social incentives in paid collaborative crowdsourcing. *ACM Trans. Intell. Syst. Technol.* **8**(6) (2017) 73:1–73:31
12. Re Calegari, G., Nasi, G., Celino, I.: Human computation vs. machine learning: an experimental comparison for image classification. *Human Computation Journal* **5**(1) (2018) 13–30
13. Siu, K., Zook, A., Riedl, M.O.: Collaboration versus competition: Design and evaluation of mechanics for games with a purpose. In: *Proceedings of Foundations of Digital Games Conference*. (2014)
14. Reeves, N., West, P., Simperl, E.: “A game without competition is hardly a game”: The impact of competitions on player activity in a human computation game. In: *Proceedings of Human Computation Conference*. (2018)
15. Reeves, N., Tinati, R., Zerr, S., Van Kleek, M., Simperl, E.: From crowd to community: A survey of online community features in citizen science projects. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017*. (2017) 2137–2152
16. Celino, I., Corcho, Ó., Hölker, F., Simperl, E.: Citizen science: Design and engagement (dagstuhl seminar 17272). *Dagstuhl Reports* **7**(7) (2017) 22–43
17. Ponciano, L., Brasileiro, F.: Finding volunteers’ engagement profiles in human computation for citizen science projects. *Human Computation Journal* **1**(2) (2015) 247–266
18. Aristeidou, M., Scanlon, E., Sharples, M.: Profiles of engagement in online communities of citizen science participation. *Computers in Human Behavior* **74** (2017) 246–256
19. Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H.R., Bertino, E., Dustdar, S.: Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing* **17**(2) (2013) 76–81
20. Karger, D.R., Oh, S., Shah, D.: Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research* **62**(1) (2014) 1–24
21. Han, T., Sun, H., Song, Y., Wang, Z., Liu, X.: Budgeted task scheduling for crowdsourced knowledge acquisition. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM* (2017) 1059–1068
22. Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? improving data quality and data mining using multiple, noisy labelers. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM* (2008) 614–622
23. Re Calegari, G., Fiano, A., Celino, I.: A Framework to build Games with a Purpose for Linked Data Refinement. In: *proceedings of the International Semantic Web Conference 2018, Resources Track*. (2018)
24. Celino, I., Re Calegari, G.: An Incremental Truth Inference Approach to Aggregate Crowdsourcing Contributions in GWAPs. In: *currently under revision*. (2018)
25. Celino, I., Contessa, S., Corubolo, M., Dell’Aglia, D., Della Valle, E., Fumeo, S., Krüger, T.: Linking Smart Cities Datasets with Human Computation: the case of UrbanMatch. In: *Proceedings of the 11th international conference on The Semantic Web, Springer-Verlag* (2012) 34–49
26. Zheng, Y., Li, G., Li, Y., Shan, C., Cheng, R.: Truth inference in crowdsourcing: is the problem solved? *Proceedings of the VLDB Endowment* **10**(5) (2017) 541–552
27. Anable, A.: Casual games, time management, and the work of affect (2013)
28. Hwang, C.: Leveling up your mobile game: Using audience measurement data to boost user acquisition and engagement (2016)