Does the Scoring Method Used in a Game Affect Learning? An Empirical Examination of a Human Resources Task

Christopher G. Harris
Department of Computer Science
SUNY Oswego
Oswego, New York 13126
christopher.harris@oswego.edu

***Abstract*—Games have been used for centuries as a tool to foster the learning process. When designed well, they have demonstrated an ability to motivate people to improve their skills and capabilities. Scoring mechanisms in games involve different strategies to improve a participant's motivation to perform well. Is the most conducive scoring method for learning providing points for good decisions, or is a scoring method that rewards points for good decisions and deducts points for poor decisions more effective? We conduct an empirical study where entry-level human resources workers (N=42) from 6 countries learn to model the behavior of experts and rate résumés on a 10-point scale. We find that a reward-only scoring mechanism benefits the learning process over one that both rewards and punishes. The reward-only scoring mechanism also provides more variability in the ratings assigned, meaning participants are more open to taking risks with their selections.**

Keywords—gamification; learning; empirical study; human resources; incentives; motivation; reward and punishment

I. Introduction

Serious games, which simultaneously provide the ability to learn new skills and be entertained, have been debated for their ability to effectively teach new skills. Many of the game-based learning efforts to date has been applied to academic settings with only a few addressing learnable decision-making scenarios that accurately reflect industry needs. This is particularly true for games that teach junior-level employees essential decision-making skills to replicate the skills of those with far more experience. Because of their subjective nature, some tasks conducted by human resources (HR) and executive search firms provide an ideal environment in which to test the transfer of decision-making and critical thinking skills from senior to junior level staff.

There is a strong need for decision-making skills in today's workplace. As the hiring needs of companies demand specialized skills, competition for talent from a limited pool of applicants becomes more acute. Selecting the most appropriate job applicants for a mid-level job position is not only subjective but also requires years of experience to accomplish well. Thus, imitating the decision-making of experts in a game-based setting, if done well, may substantially reduce the learning curve.

HR and search firms typically undertake a number of approaches to select the most appropriate résumés, such as completing a checklist or rubric for each submitted résumé and performing a keyword search on a collection of résumés. However, these approaches do not adequately address many of the nuances of a great potential employee. At the same time, the more experienced HR and executive search staff often must focus their attention on maintaining corporate accounts or attending to other needs, relegating applicant screening to junior-level employees or outsourcing it to outside firms with far less experience. This makes the task of reviewing résumés an ideal skill to teach junior HR staff using a serious game.

There are a myriad of different mechanisms to provide feedback to learners in a game-based environment, including providing badges, a score, or access to different levels [1]. Some of these feedback mechanisms involve only positive reinforcement while others involve a combination of both positive and negative reinforcement based on the performance of the participant. In this paper, we examine how these reinforcement methods, particularly scoring the inputs from the participants, might affect learning. We apply this to a task in which inexperienced participants are asked to review résumés and rate them for fit for a job description. We use a consensus of HR experts as our gold standard and measure the learning of these participants on the interrater reliability with these experts. The two research questions we examine in this study are:

1. Can gamifying a task, in which inexperienced workers are asked to replicate the decision-making of experts, show improvement after just a few sessions?
2. Is a game providing only positive scoring for making correct decisions more effective for learning quickly than a game where scoring is increased or decreased based on the decision made?

II. Background and Motivation

Ever since Gee's 2003 examination of the 36 educational principles that could be found in the design and play video games [2], serious games have emerged over the last decade as a valid mechanism to teach academic skills to students [3]. Since then, several hundred educational games and

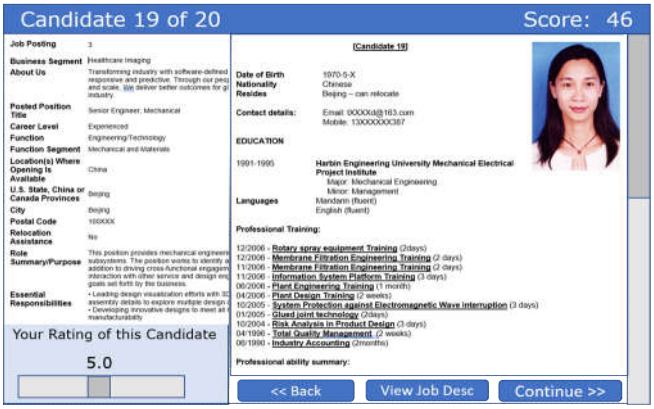

Fig. 1. Screenshot of the résumé reviewing screen, allowing the worker to provide a rating, see the candidate and the job description.

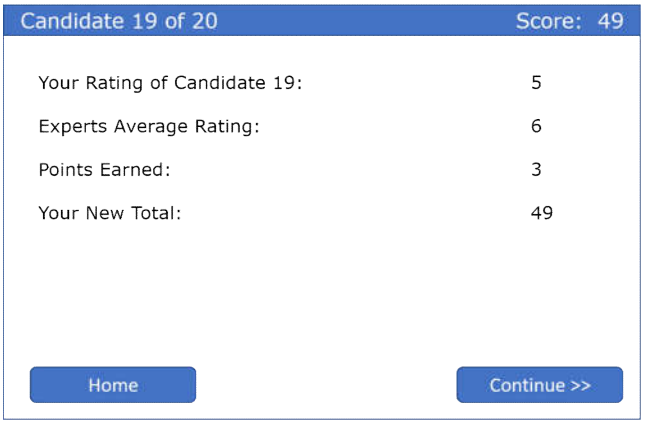

Fig. 2. A screenshot of the feedback provided to the workers in the treatment groups.

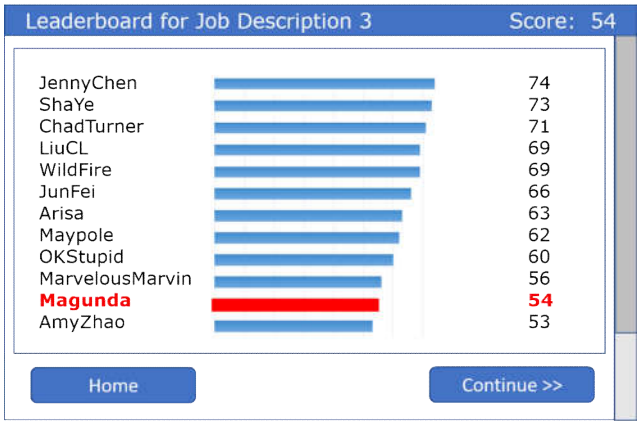

Fig. 3. A screenshot of the leaderboard for one of the job descriptions for one of the treatment groups.

simulations have been implemented to support learning across a variety of domains [4]. In response, there have been a number of studies that have examined learning benefits of serious games in many of these contexts, but the findings have ranged from a very positive impact (e.g.., [5][6] to extremely doubtful of the cognitive and motivational benefits of learning and quality outputs from serious gaming (e.g., [7][8][9]).

Recent studies by Armstrong et. al. [10] and Collmus et. al. [11] surveyed the use of games in social media (GSM) to help select and recruit participants for jobs. Another study by Lievens and Patterson [12] compared the effectiveness of low-fidelity and high-fidelity simulations for job seekers, finding that each simulation could model performance on the job.

There have been few games that have empirically investigated the learning process in HR tasks. One study by Harris [13] examined this using a extrinsic reward/punishment scheme (e.g., compensation using crowdworkers), and found that a mixed reward and punishment scoring method, where workers were rewarded for good decisions and punished for poor decisions, provided better quality than one that simply rewarded for good decisions and another that only punished for poor decisions. This study forms the foundation for our own study.

III. Experimental Methodology

A. Interface Design

Four job descriptions were taken which advertised mid-level management jobs in technical fields. To avoid potential bias, information about the hiring company was removed or made generic to avoid identifying the prospective employer. We randomly selected 4 batches of 20 résumés (80 total); one batch for each job description. These résumés were randomly selected from a pool of actual résumé submissions for each of the 4 positons. All non-standard acronyms were resolved (expanded) for clarity. All personally-identifying data were obscured or genericized to make all candidates non-identifiable. Figure 1 shows a screenshot of the résumé reviewing screen where reviews for candidates are conducted, Figure 2 shows a screenshot of the feedback given to the treatment groups, and Figure 3 shows the screenshot of the leaderboard given for each treatment and job description.

B. Obtaining the Gold Standard: Expert Ratings

Three experts in *HR* (N_{Female}=2, average number of years of experience in HR = 10.7) evaluated the résumés with respect to fit for each position. *Each expert* rated *each* résumé on an integer scale of 1 (unfit for the position) to 10 (an optimal fit). The ratings for each expert for each task is given in Table 1, indicating the experts rated the pool of candidates an average of 5.4 points, which is slightly better than average.

Table 1. Average Ratings for Each Expert, by Task

Rater	Job 1	Job 2	Job 3	Job 4	Avg. Rating
Expert 1	4.65	5.75	5.25	5.60	5.31
Expert 2	4.85	5.80	5.25	5.45	5.34
Expert 3	4.95	6.20	5.40	5.65	5.55
Average	**4.82**	**5.92**	**5.30**	**5.57**	**5.40**

We measured the inter-rater reliability between these three experts. Fleiss' Kappa (κ) ranged between 0.435 and 0.531 for the 4 tasks, averaging 0.477. The ratings are given in Table 2. This represents a "good" inter-rater reliability between these 3 expert raters [14]. It should be noted that if we had gone to a five-point rating scale instead of a ten-point

scale, the inter-rater reliability would be expected to be greater; we chose the ten-point scale (as opposed to the five-point scale used in [13]) to reduce correct selections made by chance and make the game more interesting.

TABLE 2. INTERRATER RELIBAILITY (κ) BETWEEN EXPERTS, BY JOB DESCRIPTION

	Job 1	Job 2	Job 3	Job 4	Avg
Fleiss' κ	0.435	0.454	0.487	0.531	0.477

C. Participants

In August, 2016, 42 entry-level HR workers (N_{Female} = 33, average number of years of experience in HR = 0.25) from Canada, China, Hong Kong, Malaysia, Singapore and the United States participated in our study. Each participant self-reported English abilities of an intermediate level or above. All instructions were given in English.

Each worker was asked to evaluate a single batch of 20 résumés in a single session. All workers evaluated all four batches once; batch order was randomly assigned to each worker. Workers scored each of the résumés on an integer scale of 1 to 10 with respect to the fit for the given job description for their assigned batch. Workers also had the option of scanning the entire batch of résumés before scoring them to gain familiarity with the range of applicants.

Our 42 workers were randomly assigned to either a control group or one of two treatment groups. Members of the *control group* (N=14) were not given any feedback until the end of the batch, and no points or leaderboards were provided to these participants.

The first treatment group (N=14), the *reward only treatment group*, were given the following incentive to match the average rating, rounded to the nearest integer, given by our experts: If their rating matched exactly, they were awarded 5 points; if their rating was within one point of the average expert rating, they were awarded 3 points; if their rating was within 2 points (e.g., they rated a résumé a 6 when the experts average rating was a 4), they were awarded a single point. There were no points awarded or deducted for a difference in ratings of more than two points. The scores for each worker were totaled, with a highest possible score for each batch of 100 points and the lowest score of 0 points.

A second group of workers (N=14), the *mixed reward/punishment treatment group*, were given a different incentive to match the average expert rating: much like the reward treatment group, if their rating matched exactly, they were awarded 5 points; if their rating was within one point of the average expert rating, they were awarded 3 points; if their rating was within 2, they were awarded a single point. However, if the mixed group's rating was off by more than 2, a point value, p, was assigned based on the deviation of the worker's score, w, from the expert's rating, e, using the formula $p = 5 - 2^{*} (|w - e|)$, with the maximum penalty for a single evaluation of -13 (in the case where the experts average rating for a candidate's résumé was a 10 and the worker rated it a 1, or vice versa). The scores for each worker were totaled, with a highest possible score for each batch of 100 points and the lowest score of -260 points.

D. Feedback

Each of the two treatment groups was presented with constant feedback; the average rating by each of the 3 expert raters (rounded to the nearest integer) and their own rating were given after each of the 20 résumé reviews in a batch. The control group was told they would be informally compared to the rounded average of the three experts; they received no feedback on their scoring after each résumé-job rating, but at the end of each batch of 20, they saw how their ratings compared with the HR experts.

Separate leaderboards were provided for each of the 4 batches and for each of the two treatment groups, providing a total of 8 separate leaderboards.

E. Ratings

We had 42 workers complete all 4 batches. Workers participated in a single batch each day for 4 consecutive days. Two workers did not complete the entire task of 4 batches and were replaced with other workers; the scores of these workers were removed from the study (and leaderboards).

We examined how the scores and ratings changed within the control group and two treatment groups as the workers evaluated additional batches. We wanted to see if there was some convergence in the ratings (e.g., a lazy reviewer might score every résumé a 5, which we hope to avoid), as evaluated by the standard deviation of worker's ratings as they evaluate more batches. From Table 3, we see that the control group initially had a lower first and last batch score than the treatment groups. Furthermore, the standard deviation of the ratings assigned by the control group decreased as the control group participants evaluated more batches of résumés, and was the only group to see the range of scores assigned tighten. The largest increase in standard deviation, which may imply a greater likelihood to take chances in rating candidates, was found with the reward only treatment group. Moreover, 12 of the 14 Reward Only participants had a standard deviation above 0.23 (the average standard deviation change in the control group), and all 14 participants had a positive change in their standard deviation. This may imply more confidence in rating candidates further from the median rating of 5 as they participated in more sessions.

TABLE 3. AVERAGE RATINGS AND STANDARD DEVIATIONS FOR EACH GROUP

Group	First Batch	SD First	Last Batch	SD Last	Δ Rtg	Δ SD
Control	5.15	2.31	5.13	2.08	0.02	-0.23
Reward Only	5.46	2.23	5.61	2.93	0.15	0.70
Mixed Reward/ Punish	5.38	2.18	5.63	2.40	0.25	0.22

IV. RESULTS AND ANALYSIS

Table 4 shows Fleiss' Kappa (κ), the interrater reliability of each group with the average rating of our three expert reviewers (also rounded to the nearest integer). Between the first and last batch evaluated, along with the change in κ as

more batches were evaluated. All 3 groups showed improvement as they evaluated more batches; however, the control group did not show a statistically significant increase in κ scores. There was a significant increase in the first batch κ scores for the Reward Only treatment group (M=0.2270, SD=0.0418) and last batch κ scores (M=0.4986, SD=0.0790); t(26)= 16.001, p < 0.0001 as well as for the Mixed Reward/Punishment treatment group, (M=0.2345, SD=0.0390) and last batch κ scores (M=0.4363, SD=0.0600); t(26)= 16.337, p < 0.0001. This strong contrast indicates that providing additional feedback and a game-based learning format certainly contributed to the ability for those in the treatment groups to approach the résumé scoring skills within a few days by applying a consistent gamified treatment.

TABLE 4. INTERRATER RELIABILITY (κ) FOR EACH GROUP

Group	First Batch	Last Batch	Δ IRR
Control	0.2255	0.2880	0.0625
Reward Only	0.2270	0.4986	0.2716
Mixed Reward/ Punishment	0.2345	0.4363	0.2018

Next, we assess the relative effectiveness of the scoring methods for the two treatment groups. To do this, we measure the difference in κ scores for the last batch for the two treatment groups. From Table 4, the κ score for the Reward Only group (M=0.4986, SD=0.0790) is significantly higher than the κ score for the Mixed Reward/Punishment group (M=0.4363, SD=0.0600); t(26)= 3.206, p =0.0069. This indicates the Reward Only group is slightly (but significantly) better at promoting ratings that are more consistent with the experts' consensus rating.

V. CONLCUSION

We describe a preliminary examination of learning in the context of a gamified HR task, in which novices learn to model the decision-making behavior of experts. We examined if a targeted, game-based design could improve the learning of résumé rating with entry-level HR employees from 6 countries. These employees (N=42) were divided into three groups: a control group with no gamification, one treatment group which only received points for rating job applicants close to the average consensus rating of the 3 experts, and a second treatment group that had points added or deducted from their score based on how far their rating was from the same expert consensus rating. We examined two research questions: (1) Can gamifying a task, in which inexperienced workers are asked to replicate the decision-making of experts, show improvement after just a few sessions? Also, (2) Is a game providing only positive scoring for making correct decisions more effective for learning quickly than a game where scoring is increased or decreased based on the decision made?

We learned that gamification of the résumé rating process can show very promising results after only four sessions. Also, the reward-only scoring strategy was more effective (i.e., it led to ratings that better approximated our experts) than one that rewarded participants for ratings that were close to experts and punished them for ratings that were not close. We also saw that the range of ratings became more variable (as seen by the change in the standard deviation) with the reward-only ratings system having a wider variety in ratings than in the mixed scoring system.

In future work, we anticipate looking at the longitudinal effects of the learning process over a longer time period. We also plan to increase the number of participants, examine demographics (e.g., age, gender, location), and tease out the intrinsic motivational effects have on the learning process and compare them with the extrinsic motivational effects often used as incentives in work-related tasks.

REFERENCES

[1] Mekler, E. D., Bruhlmann, F., Opwis, K., & Tuch, A. N. Do points, levels and leaderboards harm intrinsic motivation? An empirical analysis of common gamification elements. In First International Conference on Gameful Design, Research, and Applications, Toronto, Ontario, Canada, 2013 (pp. 66-73). 2583017: ACM. doi:10.1145/2583008.2583017.

[2] Gee, J. P. (2003). What video games have to teach us about learning and literacy. Computers in Entertainment (CIE), 1(1), 20-20.

[3] Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., Ruijven, T., ... & Wenzler, I. (2014). The research and evaluation of serious games: Toward a comprehensive methodology. British Journal of Educational Technology, 45(3), 502-527.

[4] Shaffer, D. W., & Gee, J. P. (2007). Epistemic games as education for innovation. Learning through digital technologies, 71-82.

[5] Wouters, P., Van Nimwegen, C., Van Oostendorp, H., & Van Der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games.

[6] Barab, S. A., Scott, B., Siyahhan, S., Goldstone, R., Ingram-Goble, A., Zuiker, S. J., & Warren, S. (2009). Transformational play as a curricular scaffold: Using videogames to support science education. Journal of Science Education and Technology, 18(4), 305.

[7] Hays, R. T. (2005). The effectiveness of instructional games: A literature review and discussion (No. NAWCTSD-TR-2005-004). Naval Air Warfare Center Training Systems Div Orlando FL.

[8] Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., ... & Yukhymenko, M. (2012). Our princess is in another castle a review of trends in serious gaming for education. Review of educational research, 82(1), 61-89.

[9] Eickhoff, C., Harris, C. G., de Vries, A. P., & Srinivasan, P. (2012). Quality through flow and immersion: gamifying crowdsourced relevance assessments. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 871-880). ACM.

[10] Armstrong, M. B., Landers, R. N., & Collmus, A. B. (2016). Gamifying recruitment, selection, training, and performance management: Game-thinking in human resource management. In Emerging research and trends in gamification (pp. 140-165). IGI Global.

[11] Collmus, A. B., Armstrong, M. B., & Landers, R. N. (2016). Game-thinking within social media to recruit and select job candidates. In Social Media in Employee Selection and Recruitment (pp. 103-124). Springer International Publishing.

[12] Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. Journal of Applied Psychology, 96(5), 927-940, doi:10.1037/a0023496.

[13] Harris, C. (2011). You're hired! an examination of crowdsourcing incentive models in human resource tasks. In Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM) (pp. 15-18).

[14] Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. Fam Med, 37(5), 360-363.