'Git Gud!' – Evaluation of Self-Rated Player Skill Compared to Actual Player Performance

Shengmei Liu, Mark Claypool
sliu7,claypool@wpi.edu
Worcester Polytechnic Institute, Worcester, MA, USA

Bhuvana Devigere, Atsuo Kuwahara, Jamie Sherman
bhuvana.devigere,atsuo.kuwahara,jamie.sherman@intel.com
Intel Corporation, Hillsboro, OG, USA

ABSTRACT

This paper evaluates the efficacy of self-rated skill as a method of differentiating player performance by analyzing data gathered in 4 previous user studies. Analysis confirms that self-rated skill can be effective for differentiating actual performance on average, but that it is not necessarily predictive for every game, and that while player performance is comparable across gender, few male participants self-rated at the lowest skill level, and no females self-rated at the highest. Additional findings suggest having participants self-rate on a five point scale, but applying those ratings in three tiers may be effective for differentiating game performance by player skill level across gender.

CCS CONCEPTS

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

KEYWORDS

skill, gamer, gender, user study, moving target

ACM Reference Format:
Shengmei Liu, Mark Claypool and Bhuvana Devigere, Atsuo Kuwahara, Jamie Sherman. 2020. 'Git Gud!' – Evaluation of Self-Rated Player Skill Compared to Actual Player Performance. In *2020 Annual Symposium on Computer-Human Interaction in Play (CHI-PLAY '20 EA), November 2–4, 2020, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3383668.3419906

1 INTRODUCTION

Most video games take time to master both in understanding what tasks meet the game challenges and in executing the tasks well. This paper grew out of a planned study which needed to ascertain player skill for recruiting participants. Past work analyzing elite gamers showed self-perceptions of skill tends to align with performance [5]. However, there is research on gender and player performance that suggests females may be under-recognized compared to males for the same skill [8]. Thus, this study was born of a practical motivation: is asking directly, as was done in these prior studies, a good way to effectively categorize players of different skill levels, also considering gender?

To answer our research questions, we use data gathered from 4 previous user studies that observed user performance for a basic game-task, with user-provided information on gender and computer gamer skill. Analysis of 181 users (25% female) playing over 700 game rounds shows:

1) Self-rated player skill is accurate in differentiating median player performance. Note, skill does not always reflect every individual performance since that can vary from game to game.

2) Self-rated skills with a 5-point scale yield only 3 tiers of differentiation: 1-2 (low), 3 (medium), and 4-5 (high). Administering self-rated skill questions on a 5 point scale, but grouping into 3 tiers in post-study analysis may help account for gender biases in the self-rating scale.

3) In our datasets, player skills are generally comparable across gender. However, only two males self-rated at the lowest skill and no females self-rated at the highest skill, despite their being no significant difference in performance between top-tier males and second-tier females.

2 DATASETS

We use 4 sets of data obtained from prior user studies [1–3]: *Mouse-A*, *Mouse-B*, *Thumbstick* and *Motion*. Each dataset was obtained from users playing a custom game with a focus on one player action – selecting a moving target with a pointing device (e.g., a mouse). Selecting a moving target is a player action common to several PC game genres (e.g., shooters).

The Mouse-A, Mouse-B and Thumbstick datasets were gathered with a custom game called Puck Hunt in which each round, the user tries to select a moving target as fast as possible. The user is scored via a timer that stops when the target is selected. Targets are 28 mm in diameter and move with three different speeds (42, 84, 126 mm/s for Mouse-A and Thumbstick and 154, 308 and 434 mm/s for Mouse-B) under 11 different added delays (0 to 400 ms), with each combination of delay and speed played 5 times.

For the first two datasets, Mouse-A and Mouse-B, users played Puck Hunt with a mouse. For the third dataset, Thumbstick, users played Puck Hunt with a game controller.

For the fourth dataset, users played a custom game called Juke! with a mouse. Juke! is like PuckHunt but the target moves with force-based physics (e.g., acceleration), with speed and direction governed by turn angle and turn frequency. Targets are 8 mm in

"Git gud" is a slang rendering of "get good", used by gamers to mean getting better at a task or skill.

CHI PLAY '20 EA, November 2–4, 2020, Virtual Event, Canada

ACM ISBN 978-1-4503-7587-0/20/11…$15.00
https://doi.org/10.1145/3383668.3419906

Table 1: Summary of dataset variables

Dataset	Users	$\overline{Age}$ (s)	Gender		Rounds	Performance	Conditions	System delay	Input
Mouse-A	51	23.7 (3.1)	43 ♂	8 ♀	167	time, clicks	3 speeds, 11 delays	20 ms	mouse
Mouse-B	31	20.9 (1.9)	23 ♂	8 ♀	167	time, clicks	3 different speeds, same delays	100 ms	mouse
Thumbstick	46	19.8 (1.5)	31 ♂	15 ♀	167	time, clicks	same as Mouse-A	50 ms	thumbstick
Motion	53	19.8 (1.5)	39 ♂	14 ♀	223	time, distance	3 turns, 3 angles, 4 delays	50 ms	mouse
Combined	181	21.1 (2.7)	136 ♂	45 ♀					

diameter and turn with an interval selected from 3 different values (30, 90, and 150 ms) and an angle from 4 different values (0, 90 and 180 degrees). The game adds a fixed amount of delay selected from 4 different values (0 to 250 ms). Each combination of jink interval, angle & delay appears 5 times.

All user studies were conducted in dedicated computer labs with computer hardware more than adequate to support the games and LCD monitors. Before playing, each participant completed informed consent forms and demographic questionnaires. The questionnaire included the question "rate yourself as a computer gamer" with responses given on a 5 point scale (1-low to 5-high). The questionnaire also included age and gender questions with options for "male", "female", "other" and "prefer not to say" – only four users did not specify either male or female. Since four is too few to provide for any meaningful statistical analysis of that group, they are removed from user counts, except where otherwise specified.

Table 1 provides a summary of the main variables in the datasets. All datasets are skewed towards young, male adults (similar to the university population drawn from) with a slight skew towards higher self-rated skill (mean slightly above 3 and mode 4 for each dataset), but there are players of all self-rated skill levels in each set.

3 ANALYSIS

Our analysis is guided by our main hypotheses that self-rated player skills correlate with player performance. Space restricts the details provided, but more information is in our technical report [10].

3.1 Player Performance

The performance of each user is the average of their target selection time across all trials in their user study. Since the games conditions are slightly different between the four studies, we normalize the data based on the average performance of all users in the same study. Users with normalized values below 1 are better than average and values above 1 are worse than average. The normalized performance values for all datasets are combined into a single dataset with one row (observation) per user: Self-rated skill (1-5), Gender (♂or ♀), and Elapsed Time (normalized seconds).

In order to assess if self-rated game skills are indicators of actual game performance, the participants' normalized selection times are grouped by self-rating of computer game skills (1-low to 5-high). A lower time is better. Figure 1 shows boxplots of normalized elapsed time on the y-axis for users clustered by self-rating on the x-axis. Each box depicts quartiles and median with the mean shown with a '+'. Points higher or lower than 1.4 × the inter-quartile range are

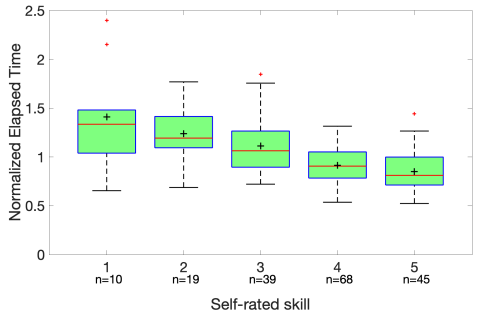

Figure 1: Elapsed time versus self-rated skill

outliers, depicted by the dots. The whiskers span from the minimum non-outlier to the maximum non-outlier. The x-axis "n=" labels indicate the number of participants in each group.

From the figure, the mean and median normalized elapsed times decrease (improve) approximately linearly with self-rated skill. However, the spread indicated by the boxes shows that some users with lower self-ratings performed better than some users with higher self-ratings.

A one-way between subjects ANOVA shows a significant effect of self-rated skill on elapsed time at the 0.05 significance level for the five conditions, $F(4, 176) = 17.86$, $p < .001$. Post-hoc tests were conducted on all self-rated skill-group pairs with corrections for multiple-comparisons. Since elapsed times were observed to be skewed right and some self-rated skill groups had fewer than 30 participants, comparisons used the Mann-Whitney U test with Bonferroni correction – effectively, testing whether two independent self-rated skill group samples come from populations having the same distribution.

Table 2 depicts the results of the Mann-Whitney U tests. Each row is a comparison between self-rated skill groups, with significant results highlighted in bold. The tests indicate that the median elapsed time is greater for skill group A than for skill group B for comparisons between 1-4, 1-5, 2-4, 2-5, 3-4, and 3-5. Median elapsed time differences between adjacent skill groups at then end of the rating scale (i.e., 1-2, 1-3, 2-3, and 4-5) are not significant.

The correlation between the elapsed times for all users and self-rated skills was significant but only weakly negatively correlated, $R^2 = 0.28$, $p < .001$. Users' predicted normalized elapsed time is equal to: $1.5 - 0.14 \times skill$, where *skill* is the self-rated skill. The correlation between the *median* elapsed time for all users and their self-rated skills was significant and strongly negatively correlated,

Table 2: Mann-Whitney U test for self-rated skill

Skill		Users		Median			
A	B	A	B	A	B	U	p value
1	2	10	19	1.32	1.18	76	0.449
1	3	10	39	1.32	1.06	117	0.055
1	**4**	10	68	1.32	0.90	105	**<.001**
1	**5**	10	45	1.32	0.81	61	**<.001**
2	3	19	39	1.18	1.06	269	0.094
2	**4**	19	68	1.18	0.90	221	**<.001**
2	**5**	19	45	1.18	0.81	110	**<.001**
3	**4**	39	68	1.06	0.90	792	**<.001**
3	**5**	39	45	1.06	0.81	404	**<.001**
4	5	68	45	0.90	0.81	1249	0.100

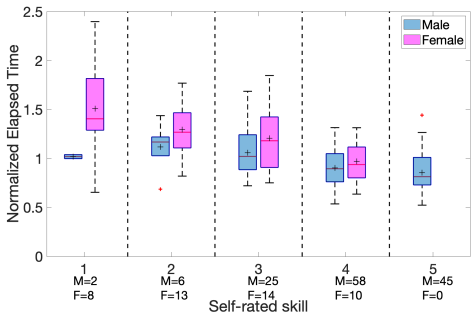

Figure 2: Elapsed time versus self-rated skill by gender

R^2 = 0.99, $p < .001$. Users' predicted median normalized elapsed time is equal to: $1.4 - 0.13 \times skill$, where *skill* is the self-rated skill.

3.2 Player Performance by Gender

Figure 2 shows boxplots as in Figure 1 but broken down by gender. The x-axis "M=" and "F=" labels indicate the number of male and female participants, respectively, in each self-rated skill group. From the figure, the mean and median elapsed times decrease approximately linearly with self-rating for both genders with the exception of males at skill 1 that has a mean and median normalized elapsed time near 1. Note, however, that there are only 2 males in this group.

A one-way between subjects ANOVA for both males and females shows a significant effect of self-rated skill on elapsed time at the 0.05 significance level for the five conditions – for males F(4, 131)) = 5.20, $p < .001$, and for females F(3, 40) = 3.78, $p = 0.018$.

The elapsed time performance of males compared to females at the same self-rated skill group were compared using Mann-Whitney U tests, the results shown in Table 3. The tests indicate differences in normalized elapsed times across genders was not significant for all skill levels.

We note that there are no females that self-rated their skills as 5, whereas 45 males (33%) self-rated their skills as 5. Visually, there is considerable overlap between the boxes for self-rated skill 4 females and self-rated skill 5 males. A Mann-Whitney U test indicates the elapsed time was not statistically different for self-rated skill 4 females (median = 0.94) than for self-rated skill 5 males (median = 0.82), U = 152, $p = 0.114$.

Table 3: Mann-Whitney U test for gender

	Users		Median			
Skill	♂	♀	♂	♀	U	p value
1	2	8	1.02	1.41	2	0.18
2	6	13	1.17	1.27	28	0.37
3	25	14	1.02	1.18	130	0.19
4	58	10	0.90	0.94	241	0.40
5	45	0	0.82	n/a	n/a	n/a

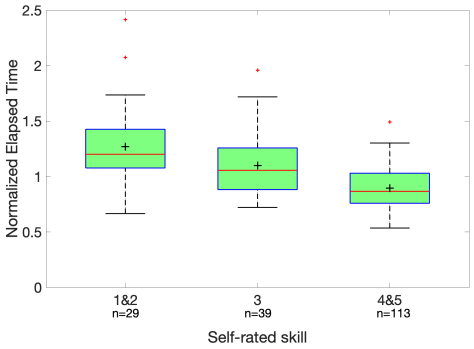

Figure 3: Elapsed time versus combined self-rated skill

3.3 Combined Self-rated Skill Groups

From Table 2, there is no statistically significant difference between the normalized elapsed times for self-rated skill groups 1 and 2 and groups 4 and 5. Hence, we explore combining self-rated skill groups 1-2 and self-rated skill groups 4-5 to effectively have 3 different skill groups: low (1 & 2), medium (3) and high (4 & 5).

Figure 3 shows boxplots as in Figure 1, but with the 1-2 and 4-5 self-rated skill groups combined. From the figure, the same visual trends hold in that mean and median normalized elapsed times decrease (improve) with self-rated skill.

A one-way between subjects ANOVA shows a significant effect of self-rated skill on elapsed time (0.05 significance) for the three conditions, F(2, 176) = 33.12, $p < .001$. Table 4 has the corresponding Mann-Whitney U tests. All results are significant, indicating the median elapsed time is greater for skill group A than for skill group B for all comparisons.

Table 4: Mann-Whitney U test for combined self-rated skill

Skill		Users		Median			
A	B	A	B	A	B	U	p value
low	**med.**	29	39	1.26	1.06	386	**0.026**
low	**high**	29	113	1.26	0.86	497	**<.001**
med.	**high**	39	113	1.06	0.86	1196	**<.001**

4 RELATED WORK

Work related to ours deals with self-efficacy ratings and facilitates a more complex discussion of how "skill" translates into execution. Dunning [4], Simons [12] and Ocay [7] point out that people tend to overestimate skills at lower levels of mastery, which might provide a way to interpret the wider variance in performance at the lowest self-rated skill. Nietfeld [6] and Ocay [7] indicate self-rated skill does not always translate globally into game success, but rather selectively in particular tasks. Shih [11] may help explain why women are more likely to self-rate at the lowest skill in seeming contradiction to the Dunning-Kruger Effect [4]. Our work builds upon this previous work by providing a specific evaluation of how well a single, self-rated gamer ability question translates into a specific game skill.

5 CONCLUSION

The goal of this research paper is to analyze the self-rating of gamer skill in relation to in-game performance. We use results from 4 previous user studies that had participants self-rate their skills and then play a game that isolated a single game action – selecting a moving target – with different game difficulties. Analysis of 181 users (136 males and 45 females) across 5 self-rated skill groups shows:

1) Self-rated skill is a strong predictor of player performance on average. For individual players, however, self-rated skill may be a weak predictor since a player's performance will vary from game to game.

2) A self-rated skill scale with 5 points only provides 3 levels of differentiation: low (self-rated scores of 1 and 2), medium (self-rated score of 3) and high (self-rated scores of 4 and 5).

3) Very few males rated themselves skill 1 and no females rated themselves skill 5. However, skills are comparable across genders - there is no significant difference between male and female performance for players that rate themselves with medium and high skill.

Our findings suggest that an effective way of differentiating player skill is to administer the self-rated skill question on a 5 point scale, but to group levels 1 and 2, and levels 4 and 5 in post-study analysis. This should allow future studies to effectively deploy player skill levels in the analysis, while helping account for gender biases in the self-rating scale.

In addition to elapsed time, player performance for targe selection can also be assessed by accuracy (number of clicks to hit the target or distance of cursor from target when clicked). Future work is to analyze such data, comparing and combining accuracy with elapsed time for a richer measure of player performance. Skill and success in games may often need more than high performance for a single game action. Thus, while we have found here that player-rated skill is predictive of performance in terms of speed, the extent to which that is a the primary driver for player success, or the most important skill for a gamer to have, remains an open research question. This includes future work assessing other game actions as well as games. Since previous research has indicated that women are less likely to identify as "gamers" [9], alternate wording for self-rated skill identification (e.g., such as "video game skill") and even multiple self-rating questions are also worth exploring.

REFERENCES

[1] Mark Claypool. 2018. Game Input with Delay - Moving Target Selection with a Game Controller Thumbstick. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) - Special Section on Delay-Sensitive Video Computing in the Cloud* 14, 3s (Aug. 2018).

[2] Mark Claypool, Andy Cockburn, and Carl Gutwin. 2019. Game Input with Delay - Moving Target Selection Parameters. In *Proceedings of the 10th Annual ACM Multimedia Systems Conference (MMSys)*. Amherst, MA, USA.

[3] Mark Claypool, Ragnhild Eg, and Kjetil Raaen. 2017. Modeling User Performance for Moving Target Selection with a Delayed Mouse. In *Proceedings of the 23rd International Conference on MultiMedia Modeling (MMM)*. Reykjavik, Iceland.

[4] D. Dunning. 2011. The Dunning-Kruger Effect: on Being Ignorant of One's Own Ignorance. *Advances in Experimental Social Psychology* 44 (2011), 247–296.

[5] Jeff Huang, Thomas Zimmermann, Nachiappan Nagappan, Charles Harrison, and Bruce Phillips. 2013. Mastering the Art of War: How Patterns of Gameplay Influence Skill in Halo. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. Paris, France.

[6] J. L. Nietfeld, L. R. Shores, and K. F. Hoffmann. 2014. Self-Regulation and Gender within a Game-Based Learning Environment. *Educational Psychology* 106, 4 (2014).

[7] A. B. Ocay. 2019. Investigating the Dunning-Kruger Effect Among Students within the Contexts of a Narrative-centered Game-based Learning Environment. In *Proceedings of the 2nd International Conference on Education Technology Management*. 8–13.

[8] Benjamin Paaßen, Thekla Morgenroth, and Michelle Stratemeyer. 2017. What is a True Gamer? The Male Gamer Stereotype and the Marginalization of Women in Video Game Culture. *Sex Roles* 76 (2017), 421 – 435.

[9] Adrienne Shaw. 2012. Do you identify as a Gamer? Gender, Race, Sexuality, and Gamer Identity. *New Media & Dociety* 14, 1 (2012), 28–44.

[10] Bhuvana Devigere Atsuo Kuwahara Jamie Sherman Shengmei Liu, Mark Claypool. 2020. *'Git Gud!' - Evaluation of Self-Rated Player Skill Compared to Actual Player Performance*. Technical Report WPI-CS-TR-20-03. Computer Science Department at Worcester Polytechnic Institute.

[11] M. Shih, T. L. Pittinsky, and N. Ambady. 1999. Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance. *Psychological Science* 10, 1 (1999), 80–83.

[12] D.J. Simons. 2013. Unskilled and Optimistic: Overconfident Predictions Despite Calibrated Knowledge of Relative Skill. *Psychonomic Bulletin & Review* 20, 3 (2013), 601–607.