WPI-CS-TR-20-03 July 2020

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

by

Shengmei Liu
Mark Claypool
Bhuvana Devigere
Atsuo Kuwahara
and Jamie Sherman

Computer Science
Technical Report
Series

WORCESTER POLYTECHNIC INSTITUTE

Computer Science Department
100 Institute Road, Worcester, Massachusetts 01609-2280

'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, Massachusetts

Bhuvana Devigere, Atsuo Kuwahara, Jamie Sherman
bhuvana.devigere,atsuo.kuwahara,jamie.sherman
Intel Corporation
Hillsboro, Oregon

ABSTRACT

It is often important to understand a player's skill level when researching the effects of delay in computer games. Past research has generally taken at face value that players' self-assessment aligns with actual abilities, yet there is also some suggestion that females may under-assess their game skills compared with males of equal ability. This paper evaluates the efficacy of self-rated skill as an effective 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 predictive for individuals, and that while player performance is generally comparable across gender, very few male participants in the collected studies rated themselves at the lowest skill level on a five point scale, and no females at all self-rated at the highest. Finally, this study found no significant difference between the performance of players in the two lowest self-rated skill tiers, and none between players in the two highest. These findings suggest that having participants self-rate on a five point scale, but applying those ratings in three tiers, may be the most effective method for differentiating actual game performance by player skill level across gender.

CCS CONCEPTS

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

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

ACM CHI PLAY '20, November 1-4, 2020, Ottawa, Ontario, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 *CHI PLAY '20: ACM CHI PLAY Conference, November 1–4, 2020, Ottawa, Ontario, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

All computer games have some delay between the moment a player provides game input until that event is processed and rendered as an image or played as a sound. These local delays can range from lows of around 20 milliseconds to highs of around 250 milliseconds [13]. Online games, including cloud-based games, have additional delays due to network processing on the end hosts and intermediate network devices [4]. Real-time games require players to execute many time-sensitive actions that degrade when delayed, and even delays less than a twentieth of a second can hamper the interplay between players' actions and their intended results. For example, delay when aiming a virtual weapon in a shooting game can make it difficult for a player to hit a moving target, decreasing the player's score and degrading the quality of experience [3].

Prior research to better understand the effects of delay on players [1, 13] has shown marked differences in the effects of delay on player performance depending upon the player's skill. Moreover, some techniques that compensate for latency [17, 19] need to consider player skill when adjusting the game world. Hence, understanding, predicting and modeling player performance requires an accurate assessment of player skill.

Most video games take some time to master both in understanding what tasks are required to meet the game challenges and in executing the tasks well. Players evaluate their competence in mastering game challenges (i.e., their skill) based on the constant, in-game performance feedback they receive while playing the game. The iterative process of mastery of

game mechanics over time and the positive feelings of control and competence over the game helps strengthen players' abilities in ex-post assessment of their own skills [16, 20]. Bandura's self-efficacy theory supports this claim and emphasizes that among numerous factors that could influence and shape self-efficacy, the most important is the experience of mastery [2]. In addition to self-efficacy theory, researchers in cognitive psychology have demonstrated one's ability to evaluate future anticipated experiences through the concept of mental models and their evolution by repeated exposure to equivalent processes and objects [10, 14]. This suggest that gamers should have a valid mental representation of their physiological and psychological conditions during gameplay, and, in fact, past work analyzing elite gamers has shown their perceptions of skill tends to align with performance [12]. However, research on gender and player performance suggests some effects of gender on performance [9] and that females may be under-recognized compared to males for the same level of skill [18]. Our personal experience with user studies and games shows that female gamers, even those with extensive experience, are much less likely to self-rate their skills at the highest levels than male gamers of similar skill and experience (see Section 3).

This paper explores the relationship between self-rated skill and actual in-game performance. Because our planned user studies have been compromised by the enforced social isolation of COVID-19 in the winter and spring of 2020, we focus instead on insights beyond a specifically targeted study and analyze data from previous studies in answering our research questions. Specifically, we use data from 4 previous user studies that observed user performance for game-tasks with delay, with user-provided information on gender and self-rated gamer skill. The goal of our analysis is to answer two main research questions:

RQ1 *Is self-rated gamer skill an effective method of estimating actual player performance?*
RQ2 *Do female players under-represent their self-rated skills?*

Analysis of 181 users (25% female) playing over 700 game rounds shows:

- Self-rated player skill is accurate in differentiating player performance on average. However, self-rated player skill does not always reflect individual player performance. These results hold for both male and female players.
- 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.
- Player skills are comparable across gender, with the exception of self-rated low skill players where males outperform females. However, across all studies, 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.

The rest of this paper is organized as follows: Section 2 summarizes previous work related to this paper, Section 3 presents the previously-conducted user studies and their resultant datasets, Section 4 describes our analysis of the results, Section 5 gives our conclusions, and Section 6 provides possible future work.

2 RELATED WORK

This section presents work related to the research questions in this paper: assessing player skill and gender and player performance.

Assessing Player Skill

Dye et al. [7] review evidence from prior studies evaluating reaction times to complete a task and corresponding accuracy. They find playing action video games significantly reduces reaction times without sacrificing accuracy, generally speeding visual tasks without decreasing performance accuracy.

Dewey [5] asserts through citations regarding music and typing that *automaticity*, gained through extensive practice, allows people to not only perform tasks quickly and accurately but also frees up cognitive resources, makes tasks take less effort over time.

Huang et al. [11] study player expertise over time by studying gameplay data from two commercial games over a 7 month time period. They find mastery of gameplay takes place through sustained and intense practice. The changes to skill take the form of bursts of improvement and can manifest as deeply ingrained, individualized habits that are available as second-nature, expert maneuvers when players are under pressure. Players who have refined their game action habits through practice have higher skills and multi-task better, particularly in time-pressured situations.

Our work studies player skills in a singular game task as it relates to their perceived (self-rated) skill.

Gender and Player Performance

Kaye and Pennington [15] examined the impact of stereotypes on female online gamers' performance. A user-study with 81 participants (majority female) show stereotype-threatened females underperform on their gaming task (collecting coins) compared to males, but social identity interventions are able

to protect females' gameplay performance. Our study directly measures female gamer performance compared to males without interventions and in relation to their self-rated skill.

Eden et al. [8] examine how two variables – game genre and player skill – inform gender perception in online games. Two experiments with 463 participants watching and rating video clips of games show that the game genre provides cues for gender perception of players – specifically, players of fighting games are perceived as more masculine than players of puzzle games. However, the perception of player skill does not provide a cue for masculinity. While their study observed relationships between player skill and perceived masculinity, our study observes relationships between self-rated skill and player performance for males and females.

Dindar [6] investigates gender differences for video game behaviors of high school students and the relationship between gaming, academic success and problem solving skills. A study of 479 participants shows that males have more gaming experience and gaming skills and spent more time playing games than the females, but there is no practical difference in academic success or problem solving skills in relation to game playing for either gender. Our study directly compares self-rated male and female gamer skills as well as their relationship between actual performance and perceived skill.

3 DATASETS

We use four sets of data obtained from prior user studies:[1] *Mouse-A*, *Mouse-B*, *Thumbstick* and *Motion*. Each dataset was obtained from users playing a game with controlled amounts of delay where the game focused on a single player action – selecting a moving target with a pointing device (e.g., a mouse). Selecting a moving target is a player action common to many PC game genres. Some examples include: 1) top-down shooters (e.g., *Nuclear Throne*, Vlambeer, 2015) where the player aims a projectile at opponents by moving the mouse to the intended target; 2) first person shooters (FPS) (e.g., *Call of Duty*, Activision, 2003) where players use the mouse to pan the game world to align a reticle over a moving opponent and shoot; and 3) multiplayer online battle arenas (MOBAs) (e.g., *League of Legends*, Riot Games, 2009) where players move a skill shot indicator with a mouse to target a moving opponent with a spell.

Games

The datasets were obtained from users playing one of two custom games: 1) Puck Hunt, used for the Mouse-A, Mouse-B, and Thumbstick datasets, and 2) Juke!, used for the Motion dataset.

[1]Citations not provided to preserve anonymity during peer-review.

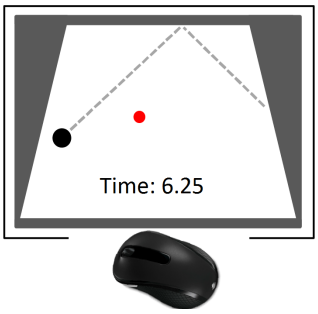

Figure 1: *Puck Hunt*. Users click on moving target (puck) with mouse cursor (red ball). Game adds delay to mouse input and varies target speeds between each round.

Puck Hunt. The Mouse-A, Mouse-B and Thumbstick datasets were gathered using a custom game called Puck Hunt that allows for the study of moving target selection with controlled amounts of delay. In Puck Hunt, depicted in Figure 1, the user proceeds through a series of short rounds, where each round has a large black ball, the puck/target, that moves with kinematics, bouncing off the edges of the screen. The user moves the mouse to control the small red ball (a.k.a., the cursor) and attempts to select the target by moving the ball over the target and clicking the mouse button. Once the user has successfully selected the target, the target disappears and a notification pops up telling the user to prepare for the next round. Thereupon pressing any key, a new round starts, with the target at a new starting location with a new orientation and speed. The user is scored via a timer that counts up from zero at the beginning of each round, stopping when the target is selected.

Users select targets, each 28 mm in diameter, with three different speeds (42, 84, 126 mm/s for the Mouse-A and Thumbstick studies and 154, 308 and 434 mm/s for the Mouse-B study) under 11 different added delays (0, 25, 50, 75, 100, 125, 150, 175, 200, 300, and 400 ms), with each combination of delay and speed encountered 5 times.

Objective measures of performance recorded are the elapsed time to select the target and the number of clicks required to do so.

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.

Juke! The fourth dataset is from a custom game called Juke!, depicted in Figure 2, that also allows study of target selection with controlled amounts of delay like Puck Hunt. However, Juke!'s target movement is governed by force-based physics (e.g., acceleration), turn angle and turn frequency. In Juke!, the user proceeds through a series of short rounds, where in each round the player moves the mouse cursor (a blue '+')

Table 1: Summary of dataset variables

Dataset	Users	$\overline{Age}$ (s)	Gender		Rounds	Performance	Conditions	System delay	Input
Set-A	51	23.7 (3.1)	43 ♂	8 ♀	167	time, clicks	3 speeds, 11 delays	20 ms	mouse
Set-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 Set-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 ♀					

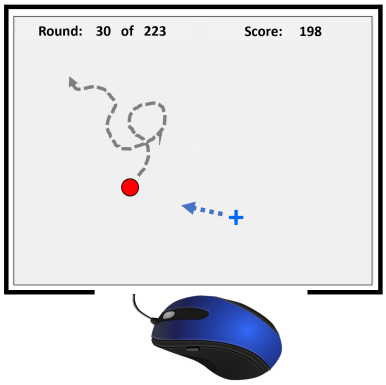

Figure 2: *Juke!* Users move the cursor (blue cross) with a mouse and click on a moving target (red circle). The game adds delay to the user input (both mouse movement and mouse clicking) and controls the target jink frequency and jink angle using force-based movement.

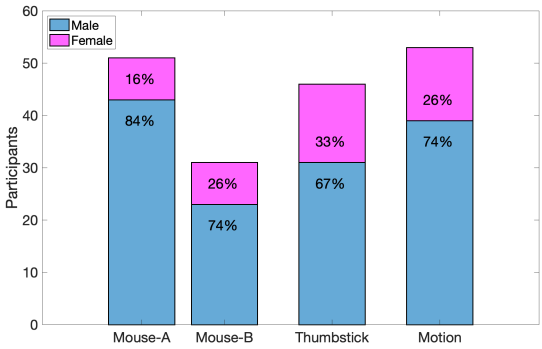

Figure 3: Dataset participants with gender breakdown

and attempts to select the target (a red ball) as quickly and accurately as possible. The user begins each round by clicking a green circle in the middle of the screen. This situates the user's mouse cursor at the same starting location each round. Upon clicking, the green circle disappears and a red target appears at a random location a short distance from the center of the screen. The user's game progress to completion is displayed in the top left corner. The user's score is displayed in the top right corner and represents a running total of the distance of the cursor from the target when clicked (lower is better).

The target, 8 mm in diameter, moves with force-based physics, applying an acceleration in the target's intended direction, with a limit on the maximum speed. The target turn interval is selected from 3 different values (30, 90, and 150 ms) and the target turn angle from 4 different values (0, 90 and 180 degrees). The game adds a fixed amount of delay selected from 4 different values (0, 62.5, 125, and 250 ms). Each combination of jink interval, angle & delay appears 5 times.

Objective measures of performance recorded are the elapsed time to click the mouse and distance between the mouse and the target when the mouse is clicked.

Procedure

All user studies were conducted in dedicated computer labs with computer hardware more than adequate to support the games and LCD monitors.

For each study, participants first completed informed consent and demographic questionnaire forms before starting the game. The demographic questionnaire included the question "rate yourself as a computer gamer" with responses given on a 5 point scale (1 - low to 5 - high). The demographic questionnaire also included an age question and a gender question with options for "male", "female", "other" and "prefer not to say" – only four users did not specify either male or female and are removed from user counts and analysis in this paper. The self-rating as a computer gamer and gender questions were mandatory.

Table 1 provides a summary of the main variables in the datasets, with the columns as follows: "Dataset" denotes the source, with the last row, "Combined" indicating the combined totals of all four datasets; the "$\overline{Age}$" column is the mean participant age in years, with the standard deviation in parentheses; "Gender" gives the breakdown of number of

males and females; "Rounds" refers to the number of game rounds the users played in Puck Hunt or Juke!; "Performance" has the user performance measures gathered by the game; "Conditions" summarizes the game configuration conditions (i.e., target motion and delay) tested; and "Input" is the user input device used by the game.

Figure 3 shows a stacked bar chart depicting the total participants in each dataset, with the blue and pink regions number (and corresponding percent) of males and females, respectively, in each dataset. Table 1 summarizes the major dataset variables, with the bottom row showing the count of users, gender and rounds of all the datasets combined into one.

Table 2 shows the breakdown of self-rated skills for each dataset, with the mean and standard deviation reported by $\bar{x}$ and s in the last two columns. The bottom row shows the breakdown of all datasets combined into one. All datasets have a slight skew towards higher self-rated skill (mean self-rated skill is slightly above 3 and the mode 4 for each dataset) but there are players of all self-rated skill levels in each set.

Table 2: Dataset breakdown of self-rated skill

	Self-rated skill						
Dataset	1	2	3	4	5	$\bar{x}$	s
Mouse-A	1	3	5	24	18	4.1	0.9
Mouse-B	4	2	9	8	8	3.5	1.3
Thumbstick	4	7	8	17	10	3.5	1.2
Motion	1	7	17	19	9	3.5	1.0
Combined	10	19	39	68	45	3.7	1.1

4 ANALYSIS

Our analysis seeks to test the following hypotheses that correspond to our research questions:

- H1 Self-rated player skills correlate with player performance
- H2 Male and female players at the same self-rated skill level perform equally well
- H3 Female players at the penultimate self-rated skill level perform as well as male players at the highest self-rated skill level

Player Performance

The performance of each user is the average of their objective measures of performance the game records (elapsed time and clicks or elapsed time and distance) across all trials in their user study. Since the games and tested conditions are slightly different between the four user studies, user results from one study cannot be directly compared (or combined) with results from another. Hence, we normalize the data for each user study based on the average performance of all users in the same study. For example, since the average elapsed time to select a target across all users and all trials for the Mouse-B dataset is 1.6 seconds, each individual user in the Mouse-B dataset has their average elapsed time divided by 1.6. Users with normalized values below 1 are better than average and values above 1 are worse than average – e.g., a normalized score of 0.9 is 10% better than the average while a 2.0 is twice as bad as the average.

The target selection games (see Section 3) have two measures of performance: a) the elapsed time to select the target, and b) the accuracy in doing so. All datasets have elapsed time, while for accuracy, the PuckHunt game (used in the Mouse-A, Mouse-B and Thumbstick datasets) has number of clicks, where each click greater than one represents a "miss", and the Juke! game (used in the Motion dataset) has the distance of the mouse from the target when clicked. In a shooting game, for example, elapsed time would measure of how quickly a player shot an opponent, distance would measure how far off the shot was, and number of clicks would measure the number of bullets used.

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 ♀), Elapsed Time (normalized seconds), and Accuracy (normalized clicks or distance).

Elapsed Time

In order to assess if self-rated game skills are indicators of actual game performance, the participants' normalized selection times are grouped by their self-ratings of computer game skills (from 1 - low to 5 - high). A lower elapsed time is better. Figure 4 shows boxplots of normalized elapsed time on the y-axis for users clustered by self-rating (1 - low to 5 - high) 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 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 that were in each self-rating 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 individuals with lower self-ratings performed better than users with higher self-ratings.

A one-way between subjects ANOVA was conducted to analyze the relationship between self-rated skill and normalized elapsed time. There was a significant effect of self-rated skill on elapsed time at the 0.05 significance level for the five conditions, $F(4, 176) = 17.86$, $p < .001$.

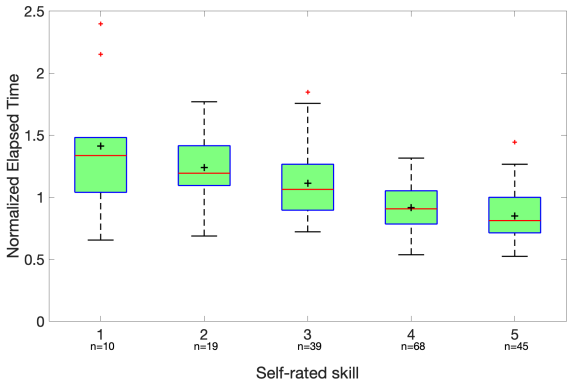

Figure 4: Elapsed time versus self-rated skill

Since the ANOVA test was statistically significant, post-hoc tests were conducted on all self-rated skill-group pairs. Since the elapsed time data was observed to be skewed right and some self-rated skill groups had fewer than 30 participants, comparisons used the Mann-Whitney U test – a non-parametric test of the null hypothesis that it is equally likely that a randomly selected value from one group will be greater than or less than a randomly selected value from a second group. Effectively, this tests whether two independent self-rated skill group samples come from populations having the same distribution.

Table 3 depicts the results of the Mann-Whitney U tests. Each row is a comparison between self-rated skill groups, labeled "A" and "B". The "Users" and "Median" columns show the number of corresponding participants and median normalized elapsed times, for the respective skill groups. The "U" and "p value" columns depict the test results. Significant results (less than 0.05) are highlighted in bold. The Mann-Whitney U 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. The tests indicates that 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 time for all users and their 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, $R^2 = 0.99$, $p < .001$. Users' predicted median normalized elapsed time is equal to: $1.4 - 0.13 \times skill$, where *skill* is the self-rated skill.

Table 3: Mann-Whitney U test for elapsed time by 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

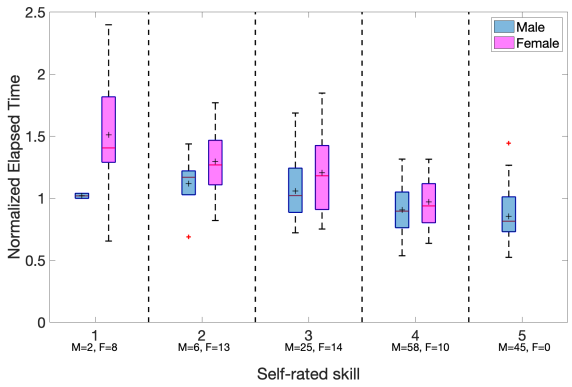

Figure 5: Elapsed time versus self-rated skill by gender

Elapsed Time Based on Gender. Figure 5 shows boxplots as in Figure 4, but with each box broken down by gender. The x-axis "M=" and "F=" labels indicate the number of male and female participants, respectively, that were 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 was conducted to analyze the relationship between self-rated skill and elapsed time for each gender. For both males and females, there was a significant effect of self-rated skill on elapsed time at the 0.05 significance level for the five conditions – for males $F(4, 131)) = 5.20$, $p < 0.001$, 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 4. Each row is a comparison between genders at the indicated self-rated skill group. The "Users" and "Median" columns show the number of corresponding participants and median normalized elapsed times for the skill groups. The "U" and "p value" columns depict the test results. The Mann-Whitney U tests indicate for all skill levels differences in normalized elapsed times across genders was not significant.

Table 4: Mann-Whitney U test for elapsed time by 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

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, $p = 0.114$.

Combined Self-rated Skill Groups

From Table 3, 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 group differentiations: low (1 & 2), medium (3) and high (4 & 5).

Figure 6 shows boxplots as in Figure 4, 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 was conducted to analyze the relationship between the combined self-rated skill groups and normalized elapsed time. There was a significant effect of self-rated skill on elapsed time at the 0.05 significance level for the three conditions, $F(2, 176) = 33.12$, $p < .001$.

Since the ANOVA test was statistically significant, post-hoc tests were conducted on the self-rated skill group combinations using Mann-Whitney U tests.

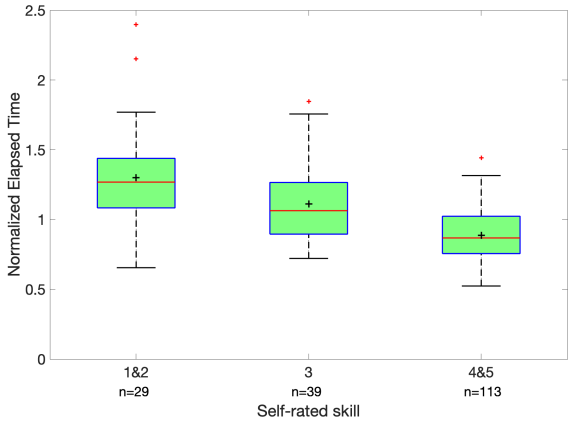

Figure 6: Elapsed time versus combined self-rated skill

Table 5: Mann-Whitney U test for elapsed time by 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**

Table 5 depicts the results of the Mann-Whitney U tests, with the rows and columns as for Table 3 except that groups 1-2 are combined into "low", 3 is "medium" and 4-5 are "high". All the results are significant (less than 0.05) so are highlighted in bold, indicating that the median elapsed time is greater for skill group A than for skill group B for all comparisons.

Figure 7 shows boxplots as in Figure 5, but with combined self-rated skill groups 1-2 and 4-5. The same visual trends of decreasing normalized elapsed time versus self-rated skill group still holds for both genders. Note, the combined self-rated skill groups provides for more of males in the lowest self-rated skill group and females in the highest self-rated skill group.

The elapsed time performance of males compared to females at the same combined self-rated skill group (low, medium, high) were compared using Mann-Whitney U tests, the results shown in Table 6. The columns are as for Table 4 and the rows have the combined self-rated skill groups for 1-2 and 4-5. The Mann-Whitney U tests indicate for medium and high skill levels differences in normalized elapsed times across genders was not significant. For the low skill group, the difference in normalized elapsed time was significant.

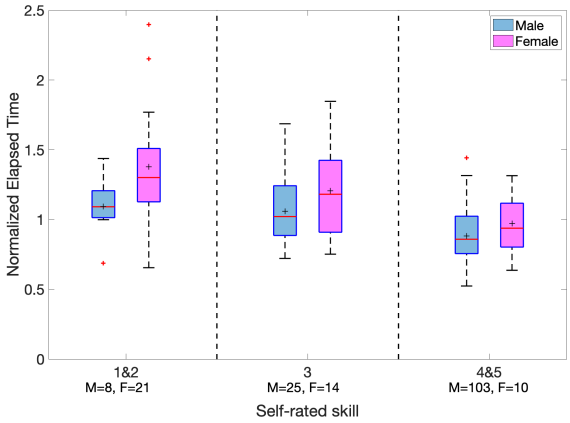

Figure 7: Elapsed time versus combined self-rated skill by gender

Table 6: Mann-Whitney U test for elapsed time by gender with combined self-rated skill

	Users		Median			
Skill	♂	♀	♂	♀	U	p value
low	8	21	1.09	1.29	42	**0.043**
med.	25	14	1.02	1.18	130	0.193
high	103	10	0.86	0.93	393	0.220

Accuracy

As indicated at the start of Section 4, in addition to elapsed time, player performance can also be assessed by accuracy (number of clicks to hit the target or distance of mouse from target when clicked). This section analyzes accuracy versus self-rated skills for the combined skill groups: low (1-2), medium (3), and high (4-5).

Figure 8 shows boxplots as in Figure 6, but the y-axis is accuracy (lower numbers are better) with the 1-2 and 4-5 self-rated skill groups combined. From the figure, the same visual trends hold for accuracy as for elapsed time in that mean and median normalized decrease (improve) with self-rated skill, although there is less separation across the medians than there is for elapsed time.

A one-way between subjects ANOVA was conducted to analyze the relationship between self-rated skill and normalized accuracy. There was a significant effect of self-rated skill on elapsed time at the 0.05 significance level for the five conditions, $F(4, 176) = 9.02$, $p < .001$. However, post-hoc tests conducted on all self-rated skill group pairs using Mann-Whitney U test showed no significant difference between groups, as seen in Table 7.

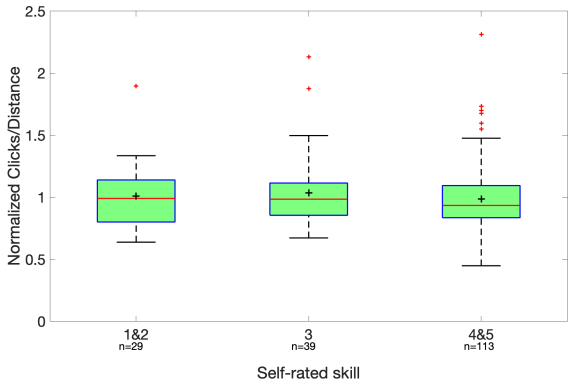

Figure 8: Accuracy versus combined self-rated skill

Figure 9 shows boxplots analysis corresponding to Figure 7, but for accuracy. From the figure, the same visual trends hold for accuracy broken down by gender, with the possible exception of males that self-rate as low skill that have a better mean accuracy than males that self-rate as medium skill. However, post-hoc tests were conducted on all self-rated skill group pairs using Mann-Whitney u test showed no significant difference between groups, as seen in Table 7.

Table 7: Mann-Whitney U test for accuracy by combined self-rated skill

Skill		Users		Median			
A	B	A	B	A	B	U	p value
low	med.	29	39	1.01	0.98	555	0.896
low	high	29	113	1.01	0.93	1474	0.411
med.	high	39	113	0.98	0.93	1949	0.286

The accuracy performance of males compared to females at the same combined self-rated skill group (low, medium, high) were compared using Mann-Whitney U tests in Table 8, with rows and columns in Table 6. The Mann-Whitney U tests indicate that for medium and high skill levels, differences in normalized accuracy across genders was not significant.

Elapsed Time versus Accuracy

There is an inherent tension between elapsed time and accuracy for target selection – i.e., in trying to be quick, a player is prone to be inaccurate, and in trying to be accurate, a player is prone to be slow. Figure 10 shows a scatter plot of normalized elapsed time versus normalized accuracy (distance/clicks). Both axes are shown in logscale (base 10) since fractional values below 1 are proportional to multiplicative values above 1 – e.g., a 10x faster elapsed time normalized

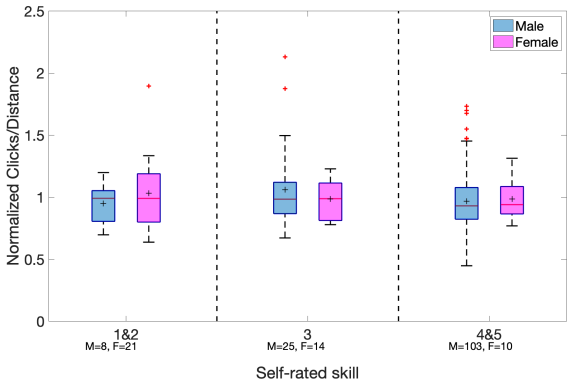

Figure 9: Accuracy versus combined self-rated skill by gender

Table 8: Mann-Whitney U test for accuracy by gender with combined self-rated skill

	Users		Median			
Skill	♂	♀	♂	♀	U	p value
low	8	21	1.00	1.02	71	0.542
med.	25	14	0.99	0.99	175	1.000
high	103	10	0.93	0.94	496	0.660

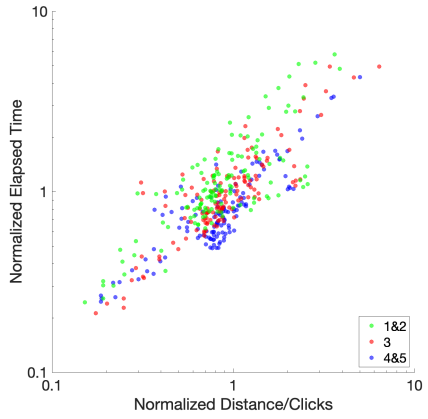

Figure 10: Normalized elapsed time versus normalized accuracy

is 0.1, which is shown the same distance away from 1 as a 10x slower elapsed time normalized to 10. Each dot is the performance of one user based on their average normalized elapsed time and average normalized accuracy across all game trials. The green dots are users with self-rated skills of 1 and 2, the red dots are users with a self-rated skill of 3, and the blue dots are users with self-rated skills of 4 and 5.

From the figure, there is a clear positive relationship between elapsed time and accuracy in that more inaccurate users tend to have higher elapsed times and vice versa. There are exceptions, however – for example, at a normalized elapsed time of 1, some users have accuracies 3x as good as the average while others have accuracies 3x as bad. Similarly, at a normalized accuracy of 1 (e.g., one click needed to select the target), some users are about twice as slow as the average while others are about twice as fast. This latter group is represented by a cluster of high-skilled users. This is supported by Dye et al. [7] who found playing action video games significantly speeds visual tasks without decreasing performance accuracy. The set of points in the bottom left of the graph make it apparent that some users have far superior performance (up to 10x better) in both elapsed time and accuracy than the average. However, these high-performance points are from all three self-rated skill groups, not just the self-rated high skilled users.

The median correlation between the log of elapsed time and log of (in)accuracy for all users was positively correlated, $R^2 = 0.66$, $p < 0.001$. Users' predicted normalized elapsed time is equal to: $0.04 + 0.77 \times accuracy$, where $accuracy$ is the normalized number of clicks or distance.

5 CONCLUSION

Knowing the skill of players that participate in user studies is important for understanding, modeling and bolstering player performance with delay. Past studies have assumed players' abilities to assess their own skills, thus informing theses studies and/or game systems tested, yet there are also indications that female players may under-represent their abilities in comparison to male players. The goal of this research paper is to analyze the self-rating of gamer skill in relation to in-game performance, both across and between genders.

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 a mouse – 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 is a weak predictor.
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 Self-rated skill is predictive of speed, but not predictive of accuracy. Higher skilled players, however, are likely to achieve the same accuracy rates as lower skilled players but in a shorter period of time.

4 Contrary to some expectations, longer elapsed times do not yield higher accuracy. Rather, slower players are also likely to be less accurate than their faster peers.
5 Skills are comparable across genders. There is no significant difference between male and female performance for medium and high skill players. Females that self-rate as low skill perform worse than males that self-rate as low skill; however, the sample size for self-rated low skill male players is extremely small.
6 Very few men rated themselves skill 1 and no women rated themselves skill 5 even though there is no statistically significant difference in the median performance of skill 4 females and skill 5 males.

These results allow us to provide answers to our research questions:

RQ1 *Is self-rated gamer skill an effective method of estimating player performance?* Answer: Sometimes. Elapsed time performance and median self-rated skill are strongly correlated. Although the individual elapsed time performance and self-rated skill correlation is weak, self-rated skill explains about 30% of the variation in performance. However, accuracy and median self-rated skill are moderately correlated and there are not statistically significant differences between self-rated skill groups.
RQ2 *Do female players under-represent their self-rated skills?* Answer: Sometimes. Female participants' self-rated skills are correlated to elapsed time and accuracy similarly to males, but high skill female participants perform as well as high skill male participants even though the males self-rate their skills higher than the females do.

These findings suggest two implications: 1) that for some atomic game actions (e.g., selecting a moving target with a mouse), pre-existing skill levels are most clearly expressed in player speed rather than accuracy, and 2) that while player self-rated skills are best differentiated along 3 tiers, a 5 point scale may be useful for normalizing skill across gender.

Regarding this second implication, given the lack of significant differentiation between the two lowest and the two highest self-rated skill groups on a 5 point scale, it is tempting to consider reducing the scale of the self-rated skill question in future studies to 3 points. However, the additional finding of this study, that female players under-represent their skills in comparison to male participants, suggests this may not be the best approach as a 3 point scale may yet again lead to female players clustering themselves in the lower tiers. Instead, our findings suggest that the most effective way of differentiating player skill may be to administer the self-rated skill question to participants on a 5 point scale, but to group levels 1 and 2, and levels 4 and 5 together in post-study analysis. Approaching self-rated skill this way will allow future studies to effectively deploy player skill levels in the analysis, while accounting for gender biases in the self-rating scale.

6 FUTURE WORK

While self-rated skill provided for differentiation in elapsed time performance, it failed to do so for accuracy. This may be because the task chosen – selecting a target – is easy enough that the "skill" part of the task is not in the player's accuracy but in the player's speed. Future work could design studies where the accuracy of the task was paramount and then ascertain if self-rated gamer skills differentiated performance.

The games in the studies used in this paper had the target appear abruptly and without warning, while most games provide anticipatory cues leading up to an event that requires players to take an action. It would be useful to understand what and how anticipatory audiovisual cues affect player performance for a given action, and how they affect sustained performance across skill levels and gender. There may emerge mitigation techniques that selectively apply anticipatory cues for players with different skill levels or genders in order to help compensate for inherent input delays, especially in network games.

Lastly, what is skill? While the player-participants provided a self-rated skill, the actual games played were novel to all. In this way, these studies and the analysis in this paper offer a perspective on aspects of gameplay that are most improved by prior experience and practice. Players who claimed higher existing skills were more likely to execute the same tasks more quickly, but not more accurately than players who professed lower skills. Yet a good number of participants who rated themselves low skill acted both faster and more quickly than some of their higher skill-rated peers. This begs the question of whether speed and accuracy are the same as skill, or even the same as performance; whether players who are faster and more accurate are more likely to perform well in mass-marketed video games. Obviously, some games are more likely to be reliant on a particular game action (e.g., target selection) than others. However, it is also worth noting that skill and success in games may often need more than high performance for a single game action, and that the focus of these studies do not account for those aspects of prior game experience and success. Some useful experiences might include familiarity with a game genre, typical interaction models, and time spent using game controllers, for example. The extent to which speed or accuracy, or something else, are the primary driver for player success, or the most important skill for a gamer to have, remains an open question

REFERENCES

[1] Rahul Amin, France Jackson, Juan E. Gilbert, Jim Martin, and Terry Shaw. 2013. Assessing the Impact of Latency and Jitter on the Perceived Quality of Call of Duty Modern Warfare 2. In *Proceedings of HCI – Users and Contexts of Use*. Las Vegas, NV, USA, 97–106.

[2] Albert Bandura. 1997. *Self-efficacy: The Exercise of Control*. W.H. Freeman, New York, NY, USA.

[3] Tom Beigbeder, Rory Coughlan, Corey Lusher, John Plunkett, Emmanuel Agu, and Mark Claypool. 2004. The Effects of Loss and Latency on User Performance in Unreal Tournament 2003. In *Proceedings of ACM Network and System Support for Games Workshop (NetGames)*. Portland, OG, USA.

[4] Kuan-Ta Chen, Yu-Chun Chang, Hwai-Jung Hsu, De-Yu Chen, Chun-Ying Huang, and Cheng-Hsin Hsu. 2014. On the Quality of Service of Cloud Gaming Systems. *IEEE Transactions on Multimedia* 26, 2 (Feb. 2014).

[5] Russell A. Dewey. 2018. *Psychology: An Introduction*. Wadsworth Publishing.

[6] Muhterem Dindar. 2018. An Empirical Study on Gender, Video Game Play, Academic Success And Complex Problem Solving Skills. *Elsevier Computers and Education* 125 (2018), 39–52.

[7] Matthew Dye, C. Shawn Green, and Daphne Bavelier. 2009. Increasing Speed of Processing With Action Video Games. *Current Directions in Psychological Science* 18, 6 (2009).

[8] Allison Eden, Erin Maloney, and Nicholas David Bowman. 2010. Gender Attribution in Online Video Games. *Journal of Media Psychology* 22 (2010), 114–124.

[9] Mona Erfani, Magy Seif El-Nasr, David Milam, Bardia Aghabeigi, Beth Aileen Lameman, Bernhard E. Riecke, Hamid Maygoli, and Sang Mah. 2010. The Effect of Age, Gender, and Previous Gaming Experience on Game Play Performance. In *Proceedings of the IFIP Human-Computer Interaction (HCI)*.

[10] W. Hacker. 1996. Action-guiding Psychological Representations ('Mental Models'). In *Encyclopedia of Psychology (in German)*, J. Kuhl and H. Heckhausen (Eds.). Gottingen Hogrefe, 769–794.

[11] Jeff Huang, Eddie Yan, Gifford Cheung, Nachiappan Nagappan, and Thomas Zimmermann. 2017. Master Maker: Understanding Gaming Skill Through Practice and Habit From Gameplay Behavior. *Topics in Cognitive Science* 9, 2 (Feb. 2017).

[12] 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.

[13] Zenja Ivkovic, Ian Stavness, Carl Gutwin, and Steven Sutcliffe. 2015. Quantifying and Mitigating the Negative Effects of Local Latencies on Aiming in 3D Shooter Games. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. Seoul, Korea, 135–144.

[14] Philip N. Johnson-Laird. 1986. *Self-efficacy: The Exercise of Control*. Harvard University Press.

[15] Linda K. Kaye and Charlotte R. Pennington. 2016. Girls Can't Play: The Effects of Stereotype Threat on Females' Gaming Performance. *Elsevier Computers in Human Behavior* 59 (2016), 202–209.

[16] Christoph Klimmt and Tilo Harmann. 2006. Effectance, Self-Efficacy and the Motivation to Play Video Games. In *Playing Video Games: Motives, Responses, and Consequences*, P. Vorderer and J. Bryant (Eds.). Lawrence Erlbaum Associates.

[17] Injung Lee, Sunjun Kim, and Byungjoo Lee. 2019. Geometrically Compensating Effect of End-to-End Latency in Moving-Target Selection Games. In *Proceedings of the ACM Computer-Human Interaction Conference (CHI)*.

[18] Benjamin PaaÃ§en, 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.

[19] Shafiee Sabet Saeed, Steven Schmidt, Saman Zadtootaghaj, Carsten Griwodz, and Sebastian Moller. 2018. Towards Applying Game Adaptation to Decrease the Impact of Delay on Quality of Experience. In *Proceedings of IEEE International Symposium on Multimedia (ISM)*. 114 – 121.

[20] Sabine Trepte and Leonard Reinecke. 2011. The Pleasures of Success: Game-Related Efficacy Experiences as a Mediator Between Player Performance and Game Enjoyment. *Cyberpsychology, Behavior, and Social Networking* 14, 9 (Sept. 2011).