Can I Pay2Win for First Person Shooting (FPS) games?

Authors: Ziyu Gan, Quentin Hsu

Introduction

E-sports have risen to be a $1 billion industry with 26.6 million monthly esports viewers in 2021. Although pro-gamers are equipped with the best of the best gear from their sponsors, the average stanford stats student is likely on a tight budget and cannot splurge on expensive equipment.

We were interested in understanding what factors can help improve an average person's skills in shooting games and whether investing in better equipment really does make a significant impact on shooting skills. In addition, we wanted to see if a stimulant like coffee will be able to further enhance a player's performance. We hypothesized that a better mouse and a higher screen refresh rate have significant effects on a player's skill but coffee does not.

We used AimLab to perform the experiment. Aim Lab is a free shooting game that players can use to train their aiming in a variety of scenarios. It outputs scores based on a combination of factors such as accuracy and speed, providing us with a standardized outcome to measure shooting performance.

Methods

Outcomes

We measured players' skill levels by the three scores that AimLab provides: overall performance score, accuracy, and number of headshots. We were primarily interested in the overall performance score to see how we could increase a player's overall FPS skill level.

Factors

The main factors we were testing included: mouse, screen refresh rate, and coffee. We were most interested in the effect of the mouse and we hypothesized that it would have the most significant effect on the player's skills. Between screen refresh rate and coffee, we were more interested in the effect of screen refresh rate. All three factors were treated as having two levels to make the experiment feasible to manage. The levels of mouse are normal mouse and gaming house. The levels of refresh rate are 60 hz and 120 hz. The levels of coffee are without coffee and without coffee.

Because we intended to collect data from multiple players over multiple games during a single gaming session, we also needed to consider the player effect (difference in the players' skills) and period effect (how many times a player has played). In other words, we treated both players and periods as additional factors in our experiment design. Because we would invite friends as players, we were more interested in each individual's skill level rather than making inferences about the overall underlying player population, we treated players as fixed effects. The **Factor Design Details** section provides more details on the levels per factor.

Experimental Design

Given the limited amount of runs we can do for our experiment (it is not feasible to run all possible combinations), we prioritized our experiment to capture the effect of mouse and screen over other potential interactions/confounders.

The following is our experiment design.

	Period							
Player	1	2	3	4	5	6	7	8
1	MS	S	M	1	1	M	S	MS
2	1	M	S	MS	MS	S	M	1
3	MS	S	M	1	1	M	S	MS
4	1	M	S	MS	MS	S	M	1
5	S	MS	1	M	M	1	MS	S
6	S	MS	1	M	M	1	MS	S
7	M	1	MS	S	S	MS	1	M
8	M	1	MS	S	S	MS	1	M

- 1 = normal mouse + 60 hz
- S = normal mouse + 120 hz
- M = gaming mouse + 60 hz
- MS = gaming mouse + 120 hz
- Brown cell = drank coffee

We got to this design by combining the strip-plot design and stepped-wedge design:

1) The player factor and time factor when crossed define the plots.

2) The 2 main factors we wanted to measure are mouse and screen. Therefore, we had 4 combinations of mouse and screen to test. Instead of splitting each plot into a subplot as in the strip-plot literature, we spread them over the plots.

3) Because we had 4 mouse and screen combinations and 8 periods, we flipped the combinations for the first 4 periods for each player to get the assignments for that player for the second 4 periods.

4) We needed to decide when we wanted to have different players to take the coffee for a stepped wedge design. We assumed a linear fixed effect model specification that included the main effects and mouse-coffee, screen-coffee, and mouse-screen interactions. We used this model specification and randomly generated 2000 coffee assignments to look for an assignment that optimized the coefficients corresponding to mouse, screen, and coffee in a descending order of priority in $(X^T X)^{-1}$. This enables us to minimize the variance of the estimators for the effect of these factors in the priority of the factors we cared about. Among the 2000 simulations, our procedure generated only a unique minimum assignment using this procedure, though in general, a minimum assignment does not need to be unique. **Appendix** provides the R code for this search procedure.

Statistical Analysis

After the data was collected, we performed linear least square estimation using the same specification that we used to optimize for coffee assignment. The Model can be written as $Y = \alpha_i + \beta_j + \gamma_k + p_t + t_m + \alpha_i\beta_j + \alpha_i\gamma_k + \beta_j\gamma_k + \epsilon$ where α, β, and γ are mouse, screen refresh rate, and coffee effects respectfully. p and t are player and time effects, each with 8 levels. ϵ is the overall noise. All three outcomes are continuous variables.

Factor Design Details

We provide the details on the mice, screen setup, and coffee that we used during the experiment.

Mouse

To simplify the mouse factor, we compared a normal wireless/bluetooth mouse to a wired gaming mouse. Gaming mice are known to have better sensors and be more accurate + responsive to use.

In particular, we had 2 normal mice and 2 gaming mice. We assumed the mice within each category were exactly the same. In reality they were slightly different generations of the same mice, but we accepted that given our limited budget.

Normal mice:

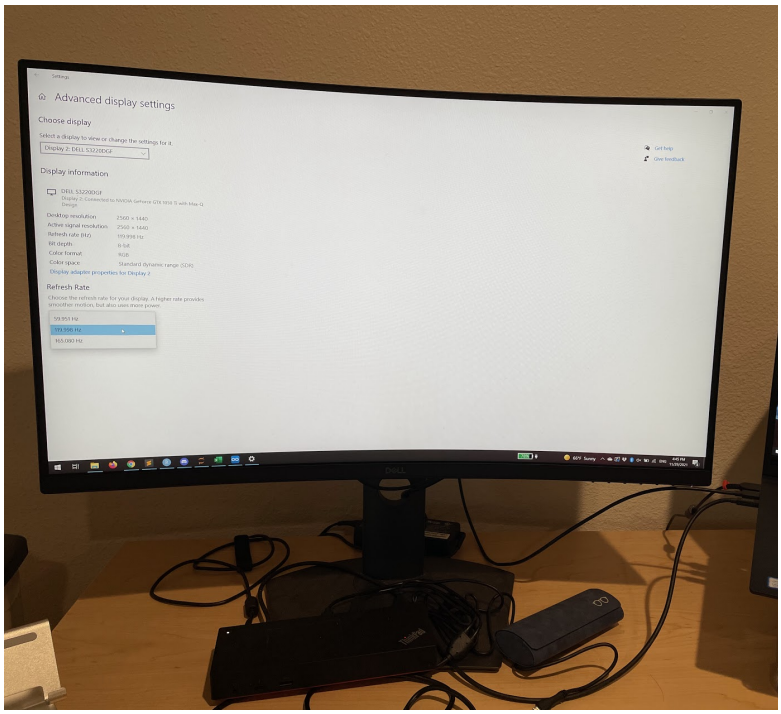
1) Logitech Performance MX
2) Logitech MX Master 3

Gaming mice:

1) Razer deathadder chroma
2) Razer deathadder elite

One aspect we could not fully test was the varying sensitivity levels used per mouse per player. Our prior knowledge suggests that sensitivity is a personalized factor with everyone comfortable at different sensitivity levels. To incorporate this personalization effect into statistical analysis, we would need to consider interaction between players and mouse sensitivity rates. We would not have enough trials if we wanted to standardize the sensitivity levels and test them with different players. As a result, we had each player find a sensitivity rate they were comfortable with for each mouse before the experiment, then kept the sensitivity levels fixed at each player's comfort level prior to each gaming round.

Screen Refresh Rate



We had 2 screens that could support both 60+ hz and 120+ hz refresh rates. Although the two screens were different models, we assumed that the effect of the model was minimal compared to screen refresh rate.

Coffee

Coffee was a particularly complicating factor as it was something that could not be turned off after turning on. As such, we opted for a split wedge design described above in order to cover enough runs to measure coffee as a factor with the mouse, screen, and time factors.
We had everyone drink 1 can of coffee after they played through their non-coffee rounds, then waited at least 15 minutes before playing again in order for the coffee effect to kick in (average length of time for coffee to take effect).

We also chose to serve cold coffee as different players would be playing at different times. Serving hot coffee would have introduced an additional variation since the temperature of the coffee would be changing depending on when a player drinks coffee.

Experiment Logistics

We planned a gaming session that had 8 participants. Since we were hosting an event, we had some responsibilities to uphold in order to be good hosts. We took into account enjoyability (our participants should enjoy the event), efficiency (our participants should only spend a reasonable amount of time for our event), and reward (we should also express our thanks to the participants) into the planning process. These responsibilities translated into the following decisions during our event planning:

- **Enjoyability** - We searched through different types of shooting tasks/games in order to find something that has a valid measure while also being enjoyable to play. Specifically, this ruled out some boring tasks like Gridshot where you just shoot and circles that appear in a grid. We eventually chose Targets of Terror, which was a Halloween exclusive event that had the player shooting at zombie figures running at them.
- **Efficiency** - We had to design our experiment to be as efficient as possible to not waste participant's time. As such, we needed to fit the experiment into a span of a lunch event and made decisions such as parallelizing the process.
- **Reward** - We catered Zareen's for participants.

Targets of Terror Snapshot

Efficient Execution

We had 64 trials to run and each trial would take around 2 minutes (setup, 1.5 minutes of task, recording scores). We also needed to wait around 15 minutes per person at different points of the experiment for the coffee effect to kick in. If we ran this naively, this could easily take over 3-4 hours. We decided to parallelize the trials with two sets of equipment.

Each player would have a first session pre-coffee, a 15-min waiting time after drinking coffee, and a second session post-coffee. As such, we broke down the trials into 8 * 3 = 24 sessions for us to organize in a parallel setup in an optimal manner. Due to the split wedge coffee design, some players had shorter pre-coffee sessions (ie. fewer periods to play in a row before drinking coffee, but more after drinking coffee) while others had shorter post-coffee sessions (ie. more periods to play in a row before drinking coffee, but fewer after drinking coffee.).

To optimize the experiment time, we took a greedy approach to select an ordering that picked shorter sessions first so that we could get many of the players drinking coffee and waiting their 15 minutes. We also clustered the longer sessions together in order to leverage our parallel setup since each session takes the longest of the 2 player sessions.

We used the following player ordering (P is for player, R is for room to play in):

Round	Room 1	Room 2	Coffee	Time
1	P3R1	P4R1	P1, P2, P3, P4	4.5 min
2	P7R1	P8R1	P7, P8	10.5 min
3	P5R1	P6R1	P5, P6	6 min
4	P1R2	P2R2		12 min
5	P3R2	P4R2		10.5 min
6	P7R2	P8R2		3 min
7	P5R2	P6R2		6 min

In the most efficient run, this plan would take less than an hour, which was very reasonable for a weekend lunch.

Results

The results we collected can be found in this Google Spreadsheet. Based on the design of our experiment, we can run a simple linear regression on all the factors/interactions we were interested in to directly get the effects of each component. Note that we set “player” and “period” as factors, which get one-hot encoded during the linear regression. All effects are fixed.

```
Call:
lm(formula = Y ~ mouse + screen + player + period + coffee +
    mouse * screen + coffee * mouse + coffee * screen, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-17073.2  -4021.0    127.6   4258.7  13834.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    91596.5     5845.3  15.670  < 2e-16 ***
mouse          15240.3     4525.6   3.368 0.001608 **
screen          8994.6     4331.9   2.076 0.043872 *
player2       -26113.6     4480.1  -5.829 6.49e-07 ***
player3       -17902.7     4528.8  -3.953 0.000283 ***
player4         2453.6     4755.2   0.516 0.608513
player5       -10572.5     4921.1  -2.148 0.037356 *
player6          558.4     4921.1   0.113 0.910192
player7       -17097.8     5456.1  -3.134 0.003106 **
player8       -18785.5     5744.1  -3.270 0.002119 **
period2         8255.4     4530.9   1.822 0.075411 .
period3        12919.0     4529.8   2.852 0.006651 **
period4        14619.6     4594.3   3.182 0.002715 **
period5        12852.6     4995.0   2.573 0.013613 *
period6        14485.6     5002.2   2.896 0.005922 **
period7        21311.2     5167.2   4.124 0.000167 ***
period8        18892.5     5422.1   3.484 0.001148 **
coffee          2080.6     5384.7   0.386 0.701114
mouse:screen  -10815.5     4513.3  -2.396 0.020980 *
mouse:coffee  -10966.9     4907.4  -2.235 0.030682 *
screen:coffee   -370.4     4953.1  -0.075 0.940734
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8960 on 43 degrees of freedom
Multiple R-squared:   0.73,     Adjusted R-squared:  0.6044
F-statistic: 5.813 on 20 and 43 DF,  p-value: 6.955e-07
```

lm results for Overall Score

The results align with our initial hypotheses.

1) Mouse has a p-value of 0.0016, which is statistically significant at a 5% level. Holding all other variables at constant, switching to the better mouse would lead to a 15240 point increase in the score. This is a relatively large effect size and could make up for the skill level difference between some of the players (eg. player4, player6). It was also comparable to the score difference from periods 4 and 6, meaning it could help make up a tiny bit of practice time.
2) Screen refresh rate has a p-value of 0.04 , which is just under the significance level of 5%. High screen refresh rate has an effect size of 8995 points. This was around half the effect size of mouse. We believe refresh rate affects players differently as those that have played shooting games before would probably notice the difference while others without a trained eye would not notice a large difference. It may also make a larger difference in real games where the accuracy requirements are much higher.
3) We see clear variation amongst the players as most of the player factors are significant. This is expected as every player's initial skill level is different. Note that based on the encoding, all the players are compared to player1.
4) Every period after the 2nd period was statistically significant. Based on the encoding, we are comparing each period to the 1st period. We see an increase in points as players play the game more. This demonstrates a "warm-up" effect.
    a) Every period has a positive increase in score compared to the 1st period, which is expected as the 1st period is the first time the players tried out the game.
    b) The 2nd period has an improvement, but not drastically since players are still getting comfortable with the game.
    c) Periods 3-6 see a significant increase, but have been relatively consistent in terms of increase in points as players go through the bulk of the games. There are a few drops and variability for the games as players switch out to drink their coffee and need to ramp up again.
    d) We see the largest amount of point increases for period 7 and 8. We heard comments from several players about trying to get their high score since they knew there were only a few games left.
5) Coffee had a p-value of 0.70 and is not statistically significant.
6) We see a negative interaction between mouse:screen and mouse:coffee. This is not expected as we thought that a better mouse and a better screen would result in higher points.

We also collected values for accuracy and headshots. Outside of player and period effects, it seems like only mouse was close to statistical significance in these cases.

```
Call:
lm(formula = accuracy ~ mouse + screen + player + period + coffee +
    mouse * screen + coffee * mouse + coffee * screen, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-15.1264  -2.5148   0.0033   2.3491   9.5826

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     65.510      3.558  18.410  < 2e-16 ***
mouse            5.499      2.755   1.996 0.052289 .
screen           0.991      2.637   0.376 0.708914
player2          8.625      2.727   3.163 0.002867 **
player3         -6.958      2.757  -2.524 0.015377 *
player4          6.221      2.895   2.149 0.037283 *
player5         -2.737      2.996  -0.914 0.366006
player6        -10.612      2.996  -3.542 0.000969 ***
player7         -7.288      3.321  -2.194 0.033668 *
player8         -3.010      3.497  -0.861 0.394051
period2          2.717      2.758   0.985 0.330006
period3          6.088      2.757   2.208 0.032627 *
period4          4.931      2.797   1.763 0.084991 .
period5          5.733      3.041   1.885 0.066141 .
period6          5.751      3.045   1.889 0.065694 .
period7          5.570      3.146   1.771 0.083664 .
period8          7.793      3.301   2.361 0.022827 *
coffee          -1.780      3.278  -0.543 0.589844
mouse:screen    -3.915      2.748  -1.425 0.161420
mouse:coffee    -2.928      2.987  -0.980 0.332586
screen:coffee    3.040      3.015   1.008 0.318919
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.455 on 43 degrees of freedom
Multiple R-squared:  0.6829,    Adjusted R-squared:  0.5354
F-statistic: 4.629 on 20 and 43 DF,  p-value: 1.267e-05
```

lm results for Accuracy (%’s)

```
Call:
lm(formula = headshots ~ mouse + screen + player + period + coffee +
    mouse * screen + coffee * mouse + coffee * screen, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-15.8121  -2.7696   0.5786   2.9085   9.3430

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     58.674      4.062  14.445  < 2e-16 ***
mouse            6.396      3.145   2.034 0.048188 *
screen          -0.418      3.010  -0.139 0.890213
player2          7.625      3.113   2.449 0.018470 *
player3         -8.788      3.147  -2.792 0.007780 **
player4          6.365      3.304   1.926 0.060718 .
player5         -4.548      3.420  -1.330 0.190564
player6        -13.423      3.420  -3.925 0.000308 ***
player7        -10.158      3.792  -2.679 0.010416 *
player8         -9.582      3.992  -2.400 0.020774 *
period2          4.322      3.149   1.373 0.176924
period3          7.639      3.148   2.427 0.019499 *
period4          6.212      3.193   1.946 0.058268 .
period5          8.489      3.471   2.446 0.018623 *
period6          7.184      3.476   2.067 0.044818 *
period7          8.211      3.591   2.287 0.027205 *
period8          9.509      3.768   2.524 0.015385 *
coffee          -2.390      3.742  -0.639 0.526347
mouse:screen    -3.052      3.136  -0.973 0.335930
mouse:coffee    -3.722      3.410  -1.091 0.281191
screen:coffee    3.811      3.442   1.107 0.274384
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.227 on 43 degrees of freedom
Multiple R-squared:  0.6945,    Adjusted R-squared:  0.5524
F-statistic: 4.887 on 20 and 43 DF,  p-value: 6.544e-06
```

lm results for Headshots (number)

Limitations

Due to a poor selection of coffee, the coffee used for the experiment had a strong bitter taste. During our experiment, we heard complaints about the taste of the coffee. Afterwards, we also noticed that some subjects did not finish the coffee that they were provided. Therefore, there were violations of both compliance and SUTVA assumptions. The coffee effect estimate may be biased downwards due to the violation of these two assumptions. Moreover, we also assumed that coffee would take effect 15 minutes after consumption and that the effect of coffee was constant over subsequent games. It may be the case that for our specific choice of coffee, the time needed for it to have effect is longer than 15 minutes and the effect varies over games. We also did not consider different reactions that people might have for coffee. Some people may need more or less coffee and time in order to feel the effect.

For the mouse factor, we only had two levels and didn't not study the effect of sensitivity. With enough budget and time, we would want to test a larger variety of mice and standardized sensitivity. It was possible that less experienced players would not know the optimal mouse sensitivity rate. Because we let the players choose a mouse sensitivity rate, the effect of mouse sensitivity rate was absorbed into the player effect. For the screen refresh rate, serious gamers would use screens whose refresh rate could reach over 240 hz. Our screens and more importantly, our computers could not support rendering at that high of a frame rate consistently. We were not able to evaluate how much gain in gaming skill level one could get by increasing screen refresh rate over 120 hz.

The true model may deviate from our model specification. When we blocked by player and by period, we assumed that different players would not interact with different periods. However, this is not necessarily true and player-period interaction could be present. For example, if a player is more experienced in FPS games, they can warm up faster. Because the coffee treatment assignment simulation was based on the assumption of a correct model specification, our estimates may be biased and coffee assignment suboptimal.

Conclusion

Player skill and practice time are necessary for getting better at shooting games. However, investing in a gaming mouse is the best return on investment compared to a better screen (which also requires a gaming computer that can run the game consistently at the higher frame rates) or drinking coffee to try to focus on the game.

Appendix

R code for obtaining coffee assignment:

```
# game plan has periods as the first 7 columns then screen, mouse,
# and 7 player columns.
mat = model.matrix(~., data=game_plan)

c_var = rep(0, 2000)
s_var = rep(0, 2000)
m_var = rep(0, 2000)
coffee_trial = rep(0, 8)

for(j in 1:2000) {
    coffee_trial = rbind(coffee_trial, coffee)
    coffee = sample(1:8, 8, replace = T)
    coffee_col = c(rep(0, 8 - coffee[1]), rep(1, coffee[1]))
    for(i in 2:8) {
        coffee_col = c(coffee_col, c(rep(0, 8 - coffee[i]), rep(1,
coffee[i])))
    }

    ms = mat[,9] * mat[, 10]
    cm =  coffee_col * mat[, 9]
    cs = coffee_col * mat[, 10]
    full_mat = cbind(mat, coffee_col)

    res = solve(t(full_mat) %*% full_mat)
    m_var[j] = res[9, 9]
    s_var[j] = res[10, 10]
    c_var[j] = res[18, 18]
}
```