BIOSTAT 681 Final Project

Cheating in Online Chess Games: A Causal Perspective

Guoxuan Ma, Kevin Christian Wibisono, Alex Liu

Dec 7, 2022

1 Introduction

We study whether there is evidence of cheating in chess matches when they are held online. There is continuous debate about whether chess players take advantage of the online environment to solicit help from outside sources that are illegal (e.g., a chess-playing bot or software) (Chess.com, 2022). We perform statistical modeling, specifically Bayesian logistic regression and propensity score matching, to provide empirical rigor to this debate.

The online and in-person chess tournament environments differ in several respects. In person, players pass through several layers of security to mitigate cheating – both metal detectors that prevent communication devices and physical building security that keeps out non-essential people limits players' interaction with the outside world. On the other hand, in the online world, cheating is detected through monitoring of the player's workspace, such as "multiple camera angles, access to screen capture, computer task manager review, audio monitoring, and sometimes even in-person on-site proctors in the room with the player" (Chess.com, 2022). However, these methods may only capture cheating after the fact, allowing faulty online tournaments to happen before cheaters can be caught. The ease of cheating in online settings is further corroborated in other domains as well, like test-taking during the COVID-19 pandemic (Newton, 2020; Lederman, 2020).

Chess tournaments earn winners tens of thousands of dollars a year. Preventing cheating in this game is important because it maintains the integrity of the game and assures that those who gain financial compensation are indeed the best players for the tournament. Understanding cheating specifically in the online context is even more urgent given that more and more games are being hosted online during the COVID-19 pandemic (Segal and McClain, 2022). This urgency to mitigate cheating is reflected in both academia (Beam, 2022) and online chess websites (Chess.com, 2022). Recently, the importance and consequences of chess cheating came to a head when Magnus Carlsen, widely considered the world's best chess player, accused Hans Niemann, an up-and-coming young star, of cheating in games both in-person and online in lucrative tournaments (Chappell, 2022). In particular, investigations into Niemann have shown that, while he likely cheated "hundreds of times" in online tournaments (a subset of which he has admitted himself), no proof exists that he cheated in person [Beaton and Robinson]. The heightened challenge of cheating in person is anecdotally evidenced by the increasingly far-flung fan theories for how Niemann cheated in person in the absence of authoritative evidence, including vibrating beads secreted inside his body (Court, 2022). At this writing, the fallout of the scandal has yet to settle.

Our proxy for cheating is upsets: instances where lower-rated players beat higher-rated players. Significant deviations in the number of upsets in online games, compared to in-person games, may be evidence of cheating. Significant deviations from expected game outcomes as a proxy for collusion or cheating has seen usage in economics literature. For instance, Duggan and Levitt (2002) found that sumo wrestlers with more to gain from a match than their opponent win at a much higher rate than their previous performance would predict them to. They took this as evidence that wrestlers collude to ”throw” a match for the needier player. Similarly, Moskowitz and Wertheim (2011) found in baseball umpires that, when the next “ball” will walk the batter then the umpire calls more strikes than usual, and when the next “strike” will get the batter out then the umpire calls more balls than usual; the author argues that this is an evidence that umpires “collude” to keep the batter in the game.

Chess tournament bodies commission investigations into the actions of players (Beam, 2022). These investigations combine qualitative methods (e.g. analyzing email exchanges and specific sets of moves) with statistical methods that estimate similarity between moves during a game with those of state-of-the-art chess engine bots. While a necessary first step to understanding the landscape of cheating, they suffer from several limitations. First, the “cheater-busing” algorithm is black-box and proprietary, and cannot be released to the public for fear of opportunistic cheaters “gaming” the detector in future matches. Second, it is individualistic: it can only analyze one player’s games at a time. Taken together, it remains an open question of whether online chess cheating occurs at a systematically higher rate as in-person chess cheating. Our study is, to the best of our knowledge, the first attempt at addressing this gap.

The rest of the report is arranged as follows. We describe details of data collection and the two methods for causal inference in Section 2. In Section 3, we provide results from both methods. We draw a conclusion and discuss the results, limitations and potential future work in Section 4.

2 Data Collection and Methodology

We first provide details on our data collection pipeline in Section 2.1. Then, we describe two methods we use for causal inference: Bayesian logistic regression in Section 2.2, and matching via propensity scores in Section 2.3.

2.1 Data Collection

We collect our data from https://chess-results.com and https://ratings.fide.com. The first website is a database containing results from more than 540,000 chess tournaments (both online and in-person) around the world. This database contains attributes of, players involved in, and outcomes of games in these tournaments. The second website provides detailed information about each chess player (as identified by their FIDE ID) such as the country they are playing for, date of birth, and rating history. We use web-scraping tools `BeautifulSoup` and `Selenium` to collect our required data from these two websites. We now describe our data collection and cleaning process, which comprises six main steps.

Step 1: We search the tournament database for tournaments whose end date lies between 2020/01/01 and 2022/10/31. Because of the COVID-19 pandemic, we expect many online tournaments to exist during this time. We observe that a majority of tournaments have low-quality data, as indicated by some games having no results and some players having no FIDE ID, among many others. In order to ensure the data we scrape are reliable, we only consider tournaments whose names contain “world” (as the website does not allow us to filter only international tournaments). We obtain 370 tournaments in total.

Step 2: We only focus on individual tournaments (not team tournaments representing companies or countries) with no missing player and game data. This brings the number of tournaments down to 201. There are 34,585 games happening in these tournaments altogether.

Step 3: We remove games in which at least one of the players has zero rating (at the time of the games).

This step is important to ensure the ratings are more representative of the players' ability. There are at least two possible explanations for a player having a rating of zero in a particular tournament: (1) inaccuracy in the database; or (2) this tournament is their first official tournament. In both cases, the zero rating most likely under-represents their true rating, which is not desirable as this rating is used to determine whether or not an upset occurs. This filtering leaves us with 27,629 games.

Step 4: We only consider games in which the difference in ratings between the two involved players are at least 200. If the difference were lower, the skill levels between these two players would be too close to consider lower-beating-higher games as real upsets. In addition, we restrict our analysis to tournaments in which both players are at least 22 years old. This is to ensure each player has enough games under their belt for their rating to be reliable. Also, this is approximately the age where most players are at or close to their peak performance. This gives us a data set comprising 2,193 games from 60 tournaments.

Step 5: For each game, we collect the following information:

- Whether the game is held online or in-person.
- If the game is held in-person, the country where the game is held.
- The start date of the tournament in which the game is held.
- The time control of the tournament (blitz, rapid or standard). This is manually done by following the rule stated here and the time increment information provided on the database.
- The FIDE ID (unique identifier) of both players.
- The ratings of both players at the tournament.
- The result of the game (win, draw or lose)

Step 6: Using these information, we generate the treatment and outcome variables, as well as the covariates for each game ("difference" here refers to higher rated player minus lower rated player):

- Binary treatment variable: medium (0 for in-person; 1 for online), denoted by A.
- Binary outcome variable: upset (0 if there is no upset; 1 if there is an upset), denoted by Y.
  - Here, upset is defined as an instance where the lower-rated player draws or wins against the higher-rated player.
- Covariate 1: time control (blitz, rapid or standard).
- Covariate 2: age difference.
- Covariate 3: home country difference.
  - We define the home country value of a player in a game to be 0 if the game is held online or in a country different from theirs; and 1 if the game is held in the same country as theirs.
  - From this definition, there are three possible values for home country difference: -1, 0 or 1.
- Covariate 4: rating difference at the tournament.
- Covariate 5: players' sex (MM, MF, FM or FF; the first letter corresponds to the higher rated player).
- Covariate 6: first game difference.
  - We define the first game value of a player in a game to be 1 if it is their first tournament in 2 years; and 0 otherwise.

  - From this definition, there are three possible values for first game difference: -1, 0 or 1.

- Covariate 7: number of tournaments difference in the past 2 years.
- Covariate 8: number of games difference in the past 2 years.
- Covariate 9: performance score difference.
  - We define the performance score of a player in a game to be the average scores (win = 1; draw = 0.5; lose = 0) of all their games in the last 2 years. If there is no game played, the performance score is set to be 0.5.
- Covariate 10: number of games in the past 2 years in which the lower rated player played against someone rated at least 200 points higher than them.
- Covariate 11: number of upsets in these games (where an upset is defined as above).
- Covariate 12: number of times the two players played against each other in the past 2 years.
- Covariate 13: number of upsets in these games (where an upset is defined as above).

Out of the 2,193 games, 2 of them are found to have incomplete covariates. Therefore, our final data set contains 2,191 games from 60 tournaments. Table 1 presents the descriptive statistics of all variables in our final data set.

2.2 Method 1: Bayesian Logistic Regression for Causal Inference

Let $\mathbf{Y}(1) = [\mathbf{Y}_{obs}^\top(1), \mathbf{Y}_{mis}^\top(1)]^\top$ and $\mathbf{Y}(0) = [\mathbf{Y}_{mis}^\top(0), \mathbf{Y}_{obs}^\top(0)]^\top$ be the binary outcomes associated with the treatment and control without loss of generality. Assume for subject i, $\mathbf{x}_i \in \mathrm{R}^{p+1}$ contains all p covariates and an intercept term. Then, $\mathbf{X} = [\mathbf{X}_1^\top, \mathbf{X}_0^\top]^\top$ where $\mathbf{X}_1 \in \mathrm{R}^{n_1 \times p+1}$ contains covariates for subjects in the treatment group and $\mathbf{X}_0 \in \mathrm{R}^{n_0 \times p+1}$ contains covariates for the control group. We specify the following Bayesian logistic regression model for estimating the causal effect for binary outcome,

$$\begin{aligned}
Y_i(1) \mid \mathbf{x}_i, \boldsymbol{\beta}_1 &\sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\boldsymbol{\beta}_1^\top \mathbf{x}_i)}\right) \\
Y_i(0) \mid \mathbf{x}_i, \boldsymbol{\beta}_0 &\sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\boldsymbol{\beta}_0^\top \mathbf{x}_i)}\right) \\
\boldsymbol{\beta}_1 \mid \sigma_1^2 &\sim \mathrm{N}(\mathbf{0}, \sigma_1^2 \mathbf{I}_{p+1}) \\
\boldsymbol{\beta}_0 \mid \sigma_0^2 &\sim \mathrm{N}(\mathbf{0}, \sigma_0^2 \mathbf{I}_{p+1}) \\
\sigma_1^2 &\sim \mathrm{IG}(a_1, b_1) \\
\sigma_0^2 &\sim \mathrm{IG}(a_0, b_0)
\end{aligned} \tag{1}$$

where a_1 and a_0 are the shape parameters, and b_1 and b_0 are the scale parameters for Inverse Gamma priors. Then, an estimate of the causal effect can be obtained by

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i(1) - Y_i(0) \Big) \tag{2}$$

where $n = n_1 + n_0$.

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
medium	2191						
... 0	1184	54%					
... 1	1007	46%					
upset	2191						
... 0	1536	70.1%					
... 1	655	29.9%					
time_control	2191						
... blitz	1250	57.1%					
... rapid	291	13.3%					
... standard	650	29.7%					
age_diff	2191	0.991	14	-53	-3	6	54
home_country_diff	2191						
... -1	230	10.5%					
... 0	1897	86.6%					
... 1	64	2.9%					
rating_diff	2191	388.77	188.113	200	249	473	1343
players_sex	2191						
... FF	305	13.9%					
... FM	94	4.3%					
... MF	267	12.2%					
... MM	1525	69.6%					
first_game_diff	2191						
... -1	112	5.1%					
... 0	1959	89.4%					
... 1	120	5.5%					
n_tour_diff	2191	1.466	13.94	-154	-5	8	58
n_game_diff	2191	16.001	125.025	-1302	-32	64	680
score_diff	2191	0.047	0.142	-0.665	-0.035	0.126	0.819
n_games_higher	2191	12.377	15.911	0	2	17	286
n_upset_higher	2191	3.957	5.394	0	0	5	75
n_peo	2191	0.165	0.61	0	0	0	7
n_upset_peo	2191	0.072	0.339	0	0	0	5

Table 1: Descriptive statistics of variables in our data set.

2.2.1 Posterior Computation for $\boldsymbol{\beta}_1$ and σ_1^2

Auxiliary variable Due to the non-conjugacy, we introduce an auxiliary random variable to develop a Gibbs sampler proposed by Polson et al. (2013). Let the auxiliary variable g_{1i} for $i = 1, 2, ..., n$ follow a Polya-Gamma distribution prior,

$$g_{1i} \sim \text{PG}(1, 0). \tag{3}$$

It can be shown that the full conditional of g_{1i} is

$$g_{1i} \mid \text{rest} \sim \text{PG}(1, \boldsymbol{\beta}_1^\top \mathbf{x}_i) \tag{4}$$

and the conditonal likelihood of $\mathbf{Y}(1)$ given $\mathbf{g}_1 = [g_{11}, g_{12}, ..., g_{1n}]^\top$ is

$$\pi(\mathbf{Y}(1) \mid \mathbf{g}_1, \text{rest}) \propto \prod_{i=1}^{n} \exp\left\{ \Big(Y_i(1) - \frac{1}{2}\Big) \boldsymbol{\beta}_1^\top \mathbf{x}_i - \frac{1}{2} g_{1i} (\boldsymbol{\beta}_1^\top \mathbf{x}_i)^2 \right\}$$

$$= \exp\left\{\sum_{i=1}^{n}\Big(Y_i(1)-\frac{1}{2}\Big)\boldsymbol{\beta}_1^\top \mathbf{x}_i - \frac{1}{2}\sum_{i=1}^{n} g_{1i}\big(\boldsymbol{\beta}_1^\top \mathbf{x}_i\big)^2\right\}$$
$$= \exp\left\{\boldsymbol{\kappa}_1^\top \mathbf{X}\boldsymbol{\beta}_1 - \frac{1}{2}\boldsymbol{\beta}_1^\top \mathbf{X}^\top \mathbf{G}_1 \mathbf{X}\boldsymbol{\beta}_1\right\} \tag{5}$$

where $\boldsymbol{\kappa}_1 = [Y_1(1)-\frac{1}{2}, Y_2(1)-\frac{1}{2}, ..., Y_n(1)-\frac{1}{2}]^\top$ and $\mathbf{G}_1 = \mathrm{diag}(g_{11}, g_{12}, ..., g_{1n})$.

Full conditional of $\boldsymbol{\beta}_1$ The full conditional of $\boldsymbol{\beta}_1$ given $\mathbf{g}_1$ is

$$\begin{aligned}\pi(\boldsymbol{\beta}_1 \mid \mathbf{g}_1, \text{rest}) &\propto \pi(\mathbf{Y}(1) \mid \mathbf{g}_1, \text{rest})\, \pi(\boldsymbol{\beta}_1 \mid \sigma_1^2)\\ &\propto \exp\left\{\boldsymbol{\kappa}_1^\top \mathbf{X}\boldsymbol{\beta}_1 - \frac{1}{2}\boldsymbol{\beta}_1^\top \mathbf{X}^\top \mathbf{G}_1 \mathbf{X}\boldsymbol{\beta}_1\right\} \exp\left\{-\frac{1}{2\sigma_1^2}\boldsymbol{\beta}_1^\top \boldsymbol{\beta}_1\right\}\\ &= \exp\left\{-\frac{1}{2}\boldsymbol{\beta}_1^\top \Big(\mathbf{X}^\top \mathbf{G}_1 \mathbf{X} + \frac{1}{\sigma_1^2}\mathbf{I}_{p+1}\Big)\boldsymbol{\beta}_1 + \boldsymbol{\kappa}_1^\top \mathbf{X}\boldsymbol{\beta}_1\right\}\\ &\propto \mathrm{N}\big(\boldsymbol{\beta}_1 \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\big)\end{aligned} \tag{6}$$

where $\boldsymbol{\Sigma}_1 = \Big(\mathbf{X}^\top \mathbf{G}_1 \mathbf{X} + \frac{1}{\sigma_1^2}\mathbf{I}_{p+1}\Big)^{-1}$ and $\boldsymbol{\mu}_1 = \boldsymbol{\Sigma}_1 \mathbf{X}^\top \boldsymbol{\kappa}_1$.

Full conditional of σ_1^2 The full conditional of σ_1^2 is

$$\sigma_1^2 \mid \text{rest} \sim \mathrm{IG}\Big(\sigma_1^2 \mid a_1 + \frac{p+1}{2},\, b_1 + \frac{1}{2}\|\boldsymbol{\beta}_1\|_2^2\Big). \tag{7}$$

2.2.2 Posterior Computation for $\boldsymbol{\beta}_0$ and σ_0^2

Auxiliary variable Similarly, let the auxiliary variable g_{0i} for $i = 1, 2, ..., n$ follow a Polya-Gamma distribution prior,

$$g_{0i} \sim \mathrm{PG}(1, 0). \tag{8}$$

It can be shown that the full conditional of g_{0i} is

$$g_{0i} \mid \text{rest} \sim \mathrm{PG}(1, \boldsymbol{\beta}_0^\top \mathbf{x}_i) \tag{9}$$

and the conditonal likelihood of $\mathbf{Y}(0)$ given $\mathbf{g}_0 = [g_{01}, g_{02}, ..., g_{0n}]^\top$ is

$$\pi(\mathbf{Y}(0) \mid \mathbf{g}_0, \text{rest}) \propto \exp\left\{\boldsymbol{\kappa}_0^\top \mathbf{X}\boldsymbol{\beta}_0 - \frac{1}{2}\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{G}_0 \mathbf{X}\boldsymbol{\beta}_0\right\} \tag{10}$$

where $\boldsymbol{\kappa}_0 = [Y_1(0)-\frac{1}{2}, Y_2(0)-\frac{1}{2}, ..., Y_n(0)-\frac{1}{2}]^\top$ and $\mathbf{G}_0 = \mathrm{diag}(g_{01}, g_{02}, ..., g_{0n})$.

Full conditional of $\boldsymbol{\beta}_0$ The full conditional of $\boldsymbol{\beta}_0$ given $\mathbf{g}_0$ is

$$\pi(\boldsymbol{\beta}_0 \mid \mathbf{g}_0, \text{rest}) \propto \mathrm{N}\big(\boldsymbol{\beta}_0 \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\big) \tag{11}$$

where $\boldsymbol{\Sigma}_0 = \Big(\mathbf{X}^\top \mathbf{G}_0 \mathbf{X} + \frac{1}{\sigma_0^2}\mathbf{I}_{p+1}\Big)^{-1}$ and $\boldsymbol{\mu}_0 = \boldsymbol{\Sigma}_0 \mathbf{X}^\top \boldsymbol{\kappa}_0$.

Full conditional of σ_0^2 The full conditional of σ_0^2 is

$$\sigma_0^2 \mid \text{rest} \sim \mathrm{IG}\Big(\sigma_0^2 \mid a_0 + \frac{p+1}{2},\, b_0 + \frac{1}{2}\|\boldsymbol{\beta}_0\|_2^2\Big). \tag{12}$$

2.2.3 Missing Data Imputation

Given the observed data and all parameters, we can estimate the missing values by the following rule,

$$\begin{aligned} Y_{mis,i}(1) &\sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\boldsymbol{\beta}_1^\top \mathbf{x}_i)}\right) \quad \text{for } i: A_i = 0 \\ Y_{mis,i}(0) &\sim \text{Bernoulli}\left(\frac{1}{1+\exp(-\boldsymbol{\beta}_0^\top \mathbf{x}_i)}\right) \quad \text{for } i: A_i = 1 \end{aligned} \tag{13}$$

2.3 Method 2: Matching via Propensity Scores

We use matching in order to estimate the average causal effect (ACE) and the average causal effect among the treated (ACT). The main goal of matching is to ensure that the distributions of subjects in the treatment and control groups are roughly identical, which is what we would expect if we had a randomized experiment. While there are many different variants of matching, we focus on propensity score matching (Rosenbaum and Rubin, 1983).

Specifically, we consider two matching methods, namely nearest neighbor (greedy) matching and full matching. For each subject on the treatment group, nearest neighbor matching selects the closest subject on the control group, where closeness is defined in terms of propensity score difference. Nearest neighbor matching can be done with or without replacement. In the latter case, the treatment observations are typically arranged in descending order of propensity scores. In full matching, each observation is assigned to one subclass consisting of one treatment and several control observations, or one control and several treatment observations (Stuart and Green, 2008). The assignment is done in a way that minimizes the sum of the absolute within-subclass distances (e.g., propensity score differences). The propensity scores are used to create the sub-classes, and the weights are computed based on subclass membership (Austin and Stuart, 2015).

We use nearest neighbor matching with and without replacement, as well as full matching to estimate the ACT, and only full matching to estimate the ACE.

3 Results and Analysis

Without accounting for the covariates, we find that the frequency of upsets is higher for in-person games as compared to online games. There are 1,184 in-person games in our sample, and 385 (32.5%) of them resulted in upsets. On the other hand, out of 1,007 online games, upsets only occurred in 270 (26.8%) of them. After adjusting for the covariates, however, we have a different story. The two subsections below explain our findings for each method in more detail.

3.1 Bayesian Logistic Regression

For prior specification, we set $a_0 = b_0 = a_1 = b_1 = 1$. We run the Gibbs sampler for 50000 steps with 25000 burn-in steps. Figure 1 shows the histogram of $\hat{\tau}$ samples and the trace plot of $\hat{\tau}$ after burn-in steps. The estimated ACE $\hat{\tau}$ by Eq. (2) is 0.0414, with a 95% Bayesian confidence interval $(0.0046, 0.0799)$, which provides significant evidence for positive ACE. We further confirm the convergence of our MCMC algorithm by checking the trace plots of σ_0^2, σ_1^2 and some elements in $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ in Figure 2.

3.2 Matching via Propensity Scores

Before estimating the ACT and ACE, it is always recommended to assess whether the matching procedure has successfully achieved balanced. We use standardized mean differences (SMDs) to quantify balance. The

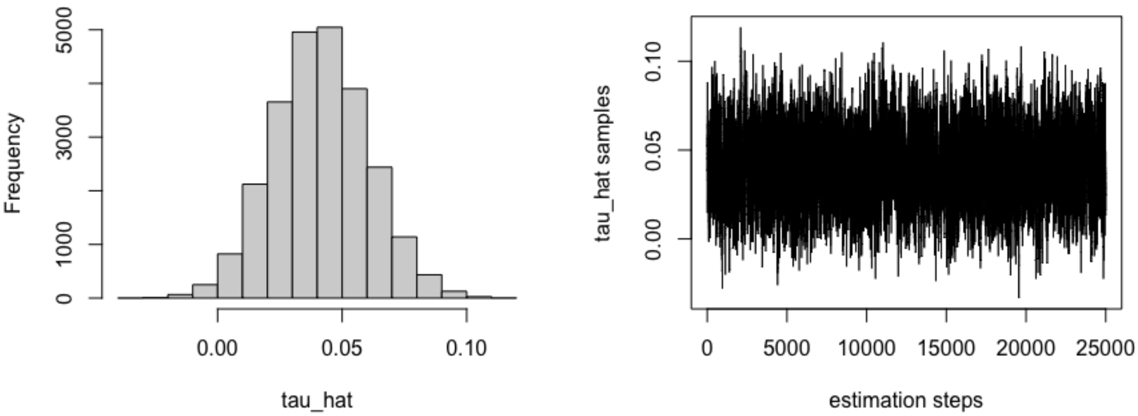


Figure 1: Left panel – histogram of $\hat{\tau}$ samples after burn-in steps; right panel – trace plot of $\hat{\tau}$ after burn-in steps.

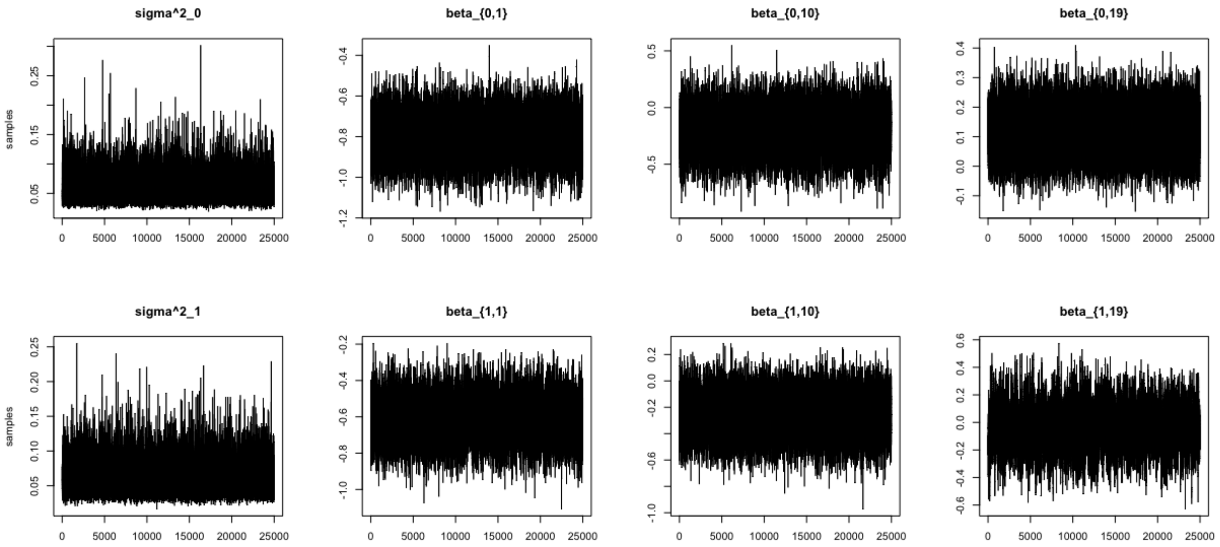
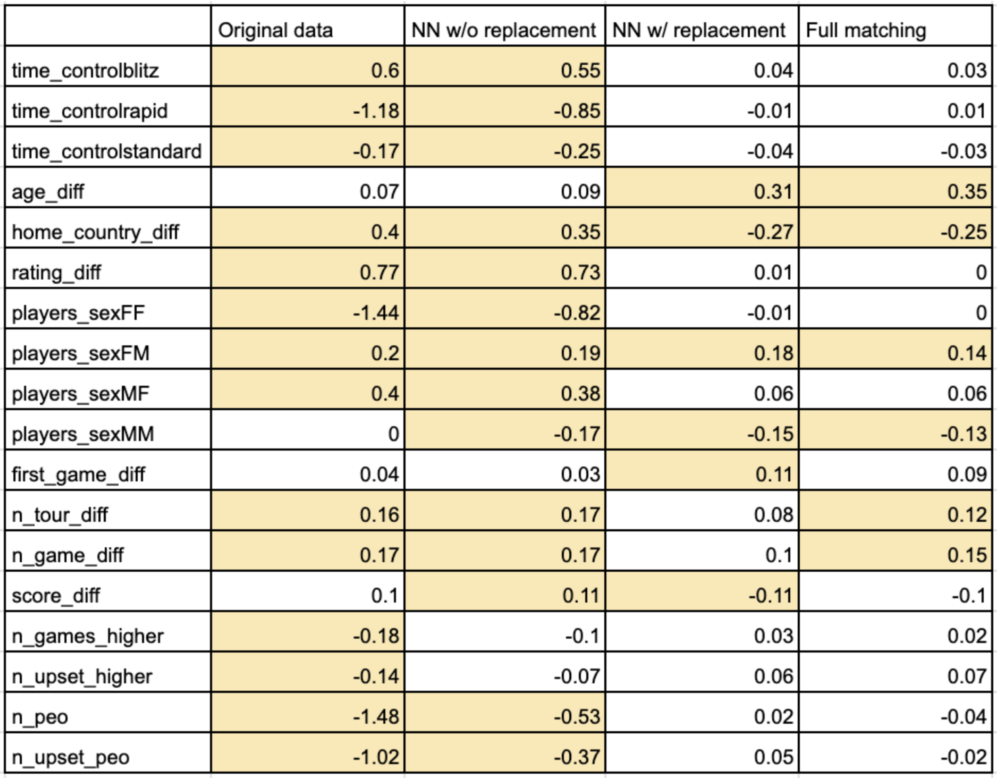


Figure 2: Trace plots of σ_0^2, σ_1^2 and some elements in $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$.

following table summarizes the SMDs for each covariate after matching using each of the three methods for estimating the ACT.

The yellow colored cells represent values which are more than 0.1 in absolute value. As seen from Table 2, nearest neighbor matching with replacement and full matching seem to work significantly better as compared to nearest neighbor matching without replacement in our data set. We therefore only consider these two methods for estimating the ACT. In order to estimate the ACT and its standard error, we fit a linear regression model with upset as the outcome; and medium, the covariates and their interactions as predictors. For full matching, we use the `comparisons` function in the `marginaleffects` package to perform g-computation, which incorporates the matching weights and cluster-robust variances. For nearest neighbor matching, since



	Original data	NN w/o replacement	NN w/ replacement	Full matching
time_controlblitz	0.6	0.55	0.04	0.03
time_controlrapid	-1.18	-0.85	-0.01	0.01
time_controlstandard	-0.17	-0.25	-0.04	-0.03
age_diff	0.07	0.09	0.31	0.35
home_country_diff	0.4	0.35	-0.27	-0.25
rating_diff	0.77	0.73	0.01	0
players_sexFF	-1.44	-0.82	-0.01	0
players_sexFM	0.2	0.19	0.18	0.14
players_sexMF	0.4	0.38	0.06	0.06
players_sexMM	0	-0.17	-0.15	-0.13
first_game_diff	0.04	0.03	0.11	0.09
n_tour_diff	0.16	0.17	0.08	0.12
n_game_diff	0.17	0.17	0.1	0.15
score_diff	0.1	0.11	-0.11	-0.1
n_games_higher	-0.18	-0.1	0.03	0.02
n_upset_higher	-0.14	-0.07	0.06	0.07
n_peo	-1.48	-0.53	0.02	-0.04
n_upset_peo	-1.02	-0.37	0.05	-0.02

Table 2: Comparison of balance among methods for estimating the ACT.

the matching is done with replacement, we need to take into account control unit multiplicity using the bootstrap (Hill and Reiter, 2006). We perform bootstraping with 1,000 bootstrap samples.

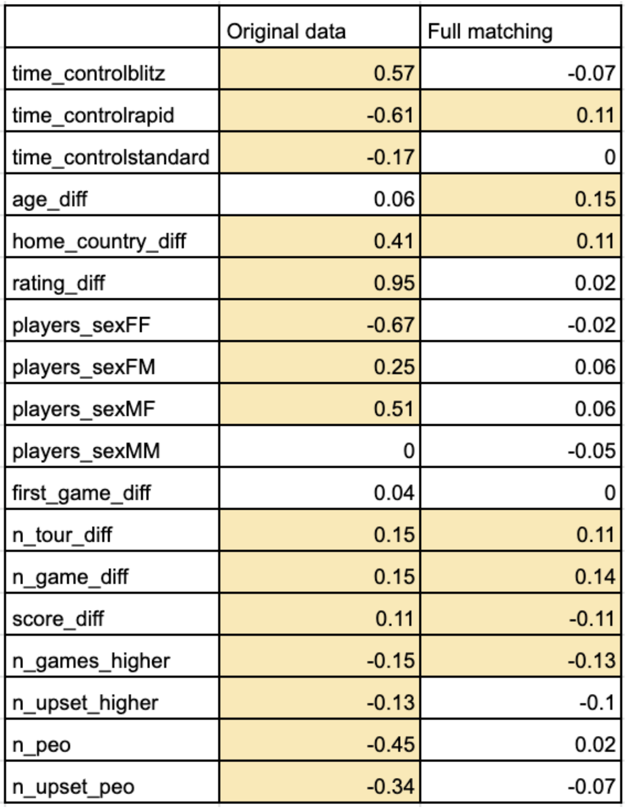
The estimated ACT from nearest neighbor matching is 0.0796, with a 95% confidence interval of $(0.0080, 0.1491)$. The estimated ACT from full matching is 0.0893, with a 95% confidence interval of $(0.0378, 0.1408)$. Both methods give evidence to support that the ACT is positive at a significance level of 0.05. To estimate the ACE and its standard error, we employ full matching after checking that it results in a reasonably good balance (see Table 3). We obtain an estimated ACE of 0.0433, with an 85% confidence interval of (0.000, 0.0860). Thus, we have evidence to support that the ACE is positive at a significance level of 0.15.

4 Conclusion, Limitations and Future Work

Overall, we find evidence that upsets occur at a significantly higher rate in online chess matches than in-person ones.

We cannot infer immediately from this finding that cheating occurs more in online contexts. The increased number of upsets may be instead attributed to other factors. For instance, playing online may be less intimidating for a lower-rated player because they are not playing in-person. Not having to see your opponent or play in front of a visible crowd may relieve them of pressure and thus boost their confidence in play. Having comfort in playing in one's home environment may also contribute to a more relaxed state for the lower-rated player. Overall, increased confidence of lower-rated players could be a post-treatment variable that we cannot account for in this experiment.

Differences in the overall ranking of players between the online and in-person group could also be a reason



	Original data	Full matching
time_controlblitz	0.57	-0.07
time_controlrapid	-0.61	0.11
time_controlstandard	-0.17	0
age_diff	0.06	0.15
home_country_diff	0.41	0.11
rating_diff	0.95	0.02
players_sexFF	-0.67	-0.02
players_sexFM	0.25	0.06
players_sexMF	0.51	0.06
players_sexMM	0	-0.05
first_game_diff	0.04	0
n_tour_diff	0.15	0.11
n_game_diff	0.15	0.14
score_diff	0.11	-0.11
n_games_higher	-0.15	-0.13
n_upset_higher	-0.13	-0.1
n_peo	-0.45	0.02
n_upset_peo	-0.34	-0.07

Table 3: Balance evaluation of full matching for estimating the ACE.

for different upset outcomes. This is because true "skill" of player does not increase linearly with the player rating. For instance, among lower-rated players, a difference of 400 points in ranking translates to a much smaller skill gap than among higher-rated players. Thus, if the online group had more lower-rated players than the in-person group, then there could be more upsets because match-ups in the online group would be between players closer in skill. To address this doubt, we plotted the mean average rating between players of a match-up, for each medium (online, in-person). We found that, while mean and median average rating was lower for online games (3113 & 3121) than offline games (3210 & 3236), these differences were not big enough to warrant any concern.

We also note that games are not independent, and that one player might appear in multiple games. Indeed, if a player won at least one game in an elimination-style tournament, they are guaranteed to have another game because they advanced to the next round of the tournament; also, if a tournament adopts the round-robin style, each player is guaranteed to play $N - 1$ times, where N is the number of players in the tournament.

In the future, we would like to develop a better proxy for cheating (e.g., inter-move times and correlation of moves with top engine moves). Moreover, more potential confounders (e.g., titles of both players) can be added in order to ensure we have a more accurate estimate of our causal effects. Lastly, a better data collection and sampling method that can result in a more diverse set of tournaments and the independence of samples will be very desirable.

References

Austin, P. C. and Stuart, E. A. (2015). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research*, 26(4):1654–1670.

Beam, C. (2022). Meet the World's Top 'Chess Detective'.

Chappell, B. (2022). The cheating scandal roiling the chess world has a new wrinkle. *NPR*.

Chess.com (2022). Hans Niemann Report.

Court, V. A. (2022). Chess champ gets butt inspected amid sex toy cheating claims.

Duggan, M. and Levitt, S. D. (2002). Winning Isn't Everything: Corruption in Sumo Wrestling. *American Economic Review*, 92(5):1594–1605.

Hill, J. and Reiter, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, 25(13):2230–2256.

Lederman, D. (2020). Best Way to Stop Cheating in Online Courses? 'Teach Better'.

Moskowitz, T. and Wertheim, L. J. (2011). *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won*. Crown. Google-Books-ID: 0lSi_eCQvNwC.

Newton, D. (2020). Another problem with shifting education online: A rise in cheating. *Washington Post*.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Segal, D. and McClain, D. L. (2022). He's the Bad Boy of Chess. But Did He Cheat? *The New York Times*.

Stuart, E. A. and Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, 44(2):395–406.