Skill vs. Chance Quantification for Popular Card & Board Games

Preprint · October 2024

DOI: 10.48550/arXiv.2410.14363

4 authors, including:

Tathagata Banerjee
Indian Institute of Technology Kanpur

Anushka De
Indian Statistical Institute

Skill vs. Chance Quantification for Popular Card & Board Games

Tathagata Banerjee[1], Anushka De[2], Subhamoy Maitra[2] & Diganta Mukherjee[2]*
[1]Indian Institute of Technology, Kanpur [2]Indian Statistical Institute, Kolkata

Abstract

We consider a few online and offline games under actual playing conditions. Generally it is expected that initially a player obtains additional skill with experience of playing more number of games and then it should finally saturate. This phase is identified when a player, with the experience of very few games, loses more when she plays against players with much longer history. Then the winning proportion curve moves up and finally it saturates. We benchmark our analysis and discussion against **Chess**, the most skilled one among the games we consider here. We use proprietary data from actual games (online and offline) as well as experiments for our statistical analysis. In this regard, we show that **Rummy** has stronger skill and learning effects. **Ludo** has similar characteristics as Rummy, but at a weaker level. Similarly, a game that is perceived as almost no skill such as **Teen Patti** indeed presents much less skill in the analysis. In the next section we describe the game structures.

1 Introduction

In this world, games serve as both entertainment and cognitive exercises, fostering strategic thinking and social interaction. Whether in the online or offline mode, they offer a diverse spectrum of experiences, engaging players in skillful pursuits and immersive narratives. Today, games are also a medium of business, drawing players from all corners of the world and impacting the economy (see [1] for survey based overview for the online Rummy game). Owing to the drastic reduction in costs of smart phones and internet usage and driven by the expansion of several interactive digital platforms, a huge market has grown for the online games among teenagers and adults.

Certain factors involved in a game could be evaluated from the player level data of the past games to resolve this question. Online games, in an Indian context, may belong to the category of the games of skill provided the success (measured, for example, by the winning rate of the same player) in such games depends on training and learning ability through experience and prowess of the players (A large literature has developed on this over the last few years, see for example [2], [3] and [4]). More specifically see [5] and [6] for analysis of digital games.

We conduct an academic study to assess the materiality of “skill” based on real user data in a few popular online card and board games. This draft analyzes the ”skill” factor associated with these games based on the winning statistics of a set of players who play against each other and have different levels of experience and capabilities [7, 3]. We take a frequentist’s approach as in [8] and showcase several elaborate results on the data gathered from a major online platform that host these games. Our Luck vs. Law modelling methodology is described in Section 3. The four games under consideration are statistically analysed in the subsequent sections sequentially. Finally, a comparative discussion is made in section 8.

*We are grateful to Dr. Tridib Mukherjee for his comments and suggestion. The usual caveat applies.

2 The Game Structures Considered

The quantification of skill vs. chance in a game has several inputs. We try to arrange these in a logical manner below.

1. Randomisation only at the beginning: we start with the unlikely example of Chess, which is considered to be a purely skill based game. The only role of chance is playing with Black or White pieces, which is statistically shown to have some implication in terms of winning ceteris paribas (discussed earlier).
   A second, and more common example is that of Rummy where the role of chance is in the initial permutation of the cards. Then a partition of which results in the two hands with the players and the sequence in the deck. We have already shown that there is significant skill (experience) premium in the Rummy game.

2. There is randomness in game play also: Ludo is a prime example where the players start with identical initial positions but nature's move (roll of the dice) dictates the choices available to them at each turn. Again, we have shown above that substantial skill premium (benefit from the use of smart strategy) exists in Ludo.

3. A second aspect is observability: whether the players can see opponent's position fully (as in Chess or Ludo) or not (in Rummy). This results into mental probability/belief updations that are refined through the course of game play. this is another aspect of Rummy that requires substantial skill.

4. Thus one may think of variations on the existing games which would highlight one or the other aspects (randomisation, observability) of a game more. Consider the Ludo Ninja variation (Zupee) which decides the rolls of the die for each turn at one go. At the very beginning a sequence of numbers in $\{1, 2, 3, 4, 5, 6\}$ are given to each player. So there is no randomisation after the start. The sequence for the opponent is unknown to a player. So there is also incomplete observability. In this sense, Ludo Ninja is closer to Rummy. This variation of Ludo also benefits from the use of smart strategy.

5. One may also think of a further variation where both the lists will be observable to the players, so that there is no lack of observability and then the game becomes one of pure strategy like Chess. Experimentation is required to understand the value of skill in this version.

So how do we proceed towards this quantification? We adopt an approach similar, but not identical, to the Getty et. al. (2018) paper in the context of fantasy games or that of [9] for Poker. Our approach will be more statistically oriented, through a regression type analysis. In this, we follow more closely our previous study on Online Rummy Game, [8]. For this purpose it is imperative to create a benchmark for Chess, which is considered closest to a pure skill game. We will use statistical methodology similar to that used for Rummy in the context of the game of Chess where skill will be proxied by ELO rating and age (in a non-monotonic manner, see [10]).

Data: For Rummy, Ludo and Teen Patti, we use proprietary data from OSG companies. For benchmarking purpose we also do extensive simulation studies. For Chess, we use publicly available game level outcome data from FIDE official website.

3 Luck vs. Law measurement formulation

The model we propose is as follows:

- Assume the following two factors are important for a favourable outcome:

- Innate skill: h, unobservable
- Learning $l_t = l(e_t)$, where e_t is experience, observable.
  Learning rate can be different between players (efficiency). But also need to consider fatigue.
- Realised Skill $s_t = s(h, l_t)$: also unobservable
- Strategy for each player: σ_1, σ_2 etc.
- Outcome (win / loss) $w_t = w(\sigma_1(s_{1t}), \sigma_2(s_{2t}), \epsilon_t) = w(\sigma_1(s(h_1, l_{1t})), \sigma_2(s(h_2, l_{2t})), \epsilon_t)$.
  Here ϵ_t is the environmental uncertainty (exogenous to the players' actions).
  The choice of strategy depending on skill creates an additional perceived uncertainty in choice of $\sigma_{1(2)}$ for player 2 (1). But this is actually part of skill.

Estimation Strategy: As skill has unobservable inputs, a direct estimation of the above relationship from (experience, outcome) data is not possible. So we posit the following:
$dw_t = \frac{\partial w}{\partial h} + \frac{\partial w}{\partial l_t}$ and $dw_{t-1} = \frac{\partial w}{\partial h} + \frac{\partial w}{\partial l_{t-1}}$. Differencing obtains $dw_t - dw_{t-1} = \frac{\partial w}{\partial l_t} - \frac{\partial w}{\partial l_{t-1}}$
As we are looking for an intuitive and easily computable benchmarking procedure, we stick to a linerised relationship. Hence, we posit the following regression equation to represent a linear approximation of the above DE:

$$w_t = \alpha + \beta_1 w_{t-1} + \beta_2 e_{t-1} + \beta_3 \Delta e_t + \text{error}$$

(here experience works as a proxy for learning).

First, consider Rummy, Ludo and TeenPatti. Consider two time intervals (0, 1) and (1, 2). suppose experience at these time points are m_i, $i = 0, 1, 2$. Then calculate $e_1 = (m_0 + m_1)/2$ and $e_2 = (m_1 + m_2)/2 - m_1 = (m_2 - m_1)/2$. Also consider the win percentage in interval (0, 1) is w_1 and in (1, 2) is w_2. For empirical convenience, we will use the following transformed version of the regression equation, that spans the whole real line:

$$\Phi^{-1}(w_2) = \alpha_0 + \alpha_1 \Phi^{-1}(w_1) + \beta_1 e_1 + \beta_2 e_2 + \text{error}$$

For Chess, we will use FIDE rating at the start of the two tournaments as e_1 and e_2.

4 Analysing Chess Tournament Data

4.1 Classical

For Classical format Chess, the leading FIDE tournament data is used. The downside of this being very small number of common players (the tournament only allows the top 14 players, see table 1). Thus results are not very reliable. In fact, none of the explanatory factors turn out to be significant (see table 2). So we continue to a version with a larger pool of common players.

Player	1st Year Win %	2nd Year Win %	ϕ^{-1}(1st Year Win %)	ϕ^{-1}(2nd Year Win %)	1st Year Rating	2nd Year Rating
Abdusattorov Nodirbek	0.65	0.61	0.39	0.29	7.27	7.13
Ding Liren	0.46	0.42	-0.096	-0.19	7.80	8.11
Giri Anish	0.65	0.65	0.39	0.39	7.49	7.64
Gukesh D	0.65	0.42	0.39	-0.19	7.25	7.25
Maghsoodloo Parham	0.35	0.54	-0.39	0.096	7.40	7.19
Praggnanandhaa R	0.58	0.46	0.19	-0.096	7.43	6.84
Van Foreest Jorden	0.35	0.46	-0.39	-0.096	6.82	6.81
L'ami Erwin	0.65	0.50	0.39	0	6.27	6.27
Roebers Eline	0.15	0.23	-1.02	-0.73	3.81	3.61
Yilmaz Mustafa	0.39	0.69	-0.29	0.502	6.65	6.09

Table 1: Player Statistics. There are only 10 common players combining Tata Masters and Challengers for Classical Chess, each have played 13 games in each year

Variable	Estimate	Std. Error	Test Statistic	p-value
Intercept	-1.1106	0.8476	-1.310	0.238
ϕ^{-1}(First Year Win Proportion)	0.1811	0.2912	0.622	0.557
First Year Rating	0.5363	0.4034	1.329	0.232
Second Year Rating	-0.3797	0.3767	-1.008	0.352

Table 2: Linear Regression Summary: Multiple $R^2 = 0.483$

4.2 Rapid, one year data

In the version of Chess, there is a broader format that allows a larger number of players to play against each other. Now we first use only one year, a single edition of the tournament data first. Since no. of games is < 13 for some players (maybe they quit after some rounds), took both half game count as predictor. Dropped second half as it's not significant. See the result in table 3. Now the explanatory factors return significant values with ratings having a positive effect but performance in the first half becoming negative. This could be due to the correlation between performance and rating. To understand this better, we move to a two year analysis next.

Variable	Estimate	Std. Error	Test Statistic	p-value
Intercept	-3.08	0.47	-6.51	0
ϕ^{-1}(First Half Win Proportion)	-0.36	0.08	-4.51	0
Rating	0.16	0.021	7.52	0
First Half Games Count	0.29	0.06	4.65	0

Table 3: Regression Summary: Multiple $R^2 = 0.266$.

4.3 Rapid, two year data

In this extended version of the analysis using two years' data, we have a total of 96 players common in both editions of the tournament. Initial and final ratings denote ratings in 2022 and 2023 respectively. Response is ϕ^{-1}(win proportion in 23). As one can see from the result in table 4, now performance, initial rating and second year rating all has the correct sign and are significant.

Variable	**Estimate**	**Std. Error**	**Test Statistic**	**p-value**
Intercept	-1.09	0.14	7.99	0
Initial Rating	0.34	0.06	5.68	0
Final Rating	0.17	0.05	3.39	0.001
ϕ^{-1}(Initial Win Proportion)	0.12	0.09	1.41	0.163

Table 4: Regression Model Summary: $R^2 = 0.611$

4.4 Chess Online Data

For an empirical investigation of the robustness of our findings, we need a tournament setting where a large pool of players play against each other. These results can be used to empirically define the variability (or otherwise) of our results using resampling techniques.

4.4.1 Resampling Procedure

We have applied the idea of empirical bootstrap to the regression problem. In this case, the empirical bootstrap is also called paired bootstrap. Given the original sample $(X_1, Y_1), \cdots, (X_n, Y_n)$, we generate a new sets of IID observations

$$\left(X_1^*, Y_1^*\right), \cdots, \left(X_n^*, Y_n^*\right)$$

such that for each ℓ,

$$P\left(X_\ell^* = X_i, Y_\ell^* = Y_i\right) = \frac{1}{n}, \quad \forall i = 1, \cdots, n$$

Namely, we treat (X_i, Y_i) as one object and we sample with replacement n times from these n objects to form a new bootstrap sample. Thus, each time we generate a set of n new observations from the original dataset. Assume we repeat the entire process B times, we would obtain

$$\begin{gathered}\left(X_1^{*(1)}, Y_1^{*(1)}\right), \cdots, \left(X_n^{*(1)}, Y_n^{*(1)}\right) \\ \left(X_1^{*(2)}, Y_1^{*(2)}\right), \cdots, \left(X_n^{*(2)}, Y_n^{*(2)}\right) \\ \vdots \\ \left(X_1^{*(B)}, Y_1^{*(B)}\right), \cdots, \left(X_n^{*(B)}, Y_n^{*(B)}\right)\end{gathered}$$

For each bootstrap sample, say $\left(X_1^{*(\ell)}, Y_1^{*(\ell)}\right), \cdots, \left(X_n^{*(\ell)}, Y_n^{*(\ell)}\right)$, we fit the linear regression, leading to a bootstrap estimate of the fitted coefficients $\widehat{\beta}_0^{*(\ell)}, \widehat{\beta}_1^{*(\ell)}$. Thus, the B bootstrap samples leads to

$$\left(\widehat{\beta}_0^{*(1)}, \widehat{\beta}_1^{*(1)}\right), \cdots, \left(\widehat{\beta}_0^{*(B)}, \widehat{\beta}_1^{*(B)}\right)$$

B sets of fitted coefficients. We then estimate the variance by

$$\widehat{\mathrm{Var}}_B\left(\widehat{\beta}_0\right) = \frac{1}{B}\sum_{\ell=1}^{B}\left(\widehat{\beta}_0^{*(\ell)} - \bar{\beta}_0^*\right)^2, \quad \bar{\beta}_0^* = \frac{1}{B}\sum_{\ell=1}^{B}\widehat{\beta}_0^{*(\ell)}$$

$$\widehat{\mathrm{Var}}_B\left(\widehat{\beta}_1\right) = \frac{1}{B}\sum_{\ell=1}^{B}\left(\widehat{\beta}_1^{*(\ell)} - \bar{\beta}_1^*\right)^2, \quad \bar{\beta}_1^* = \frac{1}{B}\sum_{\ell=1}^{B}\widehat{\beta}_1^{*(\ell)}$$

We can construct the confidence intervals using the variance estimate:

$$\text{C.I. } (\beta_0) = \widehat{\beta}_0 \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathrm{Var}}_B\left(\widehat{\beta}_0\right)}$$

$$\text{C.I. } (\beta_1) = \widehat{\beta}_1 \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathrm{Var}}_B\left(\widehat{\beta}_1\right)}$$

4.4.2 Regression Results

For this study, data has been collected from Lichess, an online chess platform. Data of blitz games of March and April 2024 have been used, and only games played by players rated 2300+ on Lichess have been considered. Further, only the players having played at least 5 games in each month have been considered for the study. The database consists of data for 5267 such players.

Since Lichess rating is updated after every game(unlike FIDE tournaments, where rating is not updated between tournament games), the experience gained in the months of March and April (considered to be the two halves) have been quantified in the following way-

- *First Half Experience* = $\frac{m_0+m_1}{2}$, where m_0 and m_1 are the ratings after the first and last matches in March respectively.
- *Second Half Experience* = $\frac{m_2+m_3}{2} - m_1$, where m_2 and m_3 are the ratings after the first and last matches in April respectively. (Note: players might have unrecorded matches between the calculation of m_1 and m_2, since the data only contains games where both players are rated 2300+ .)

It is to be noted that Lichess ratings and FIDE ratings are calculated slightly differently, and we perform origin-scale transformation on the experiences before using them for regression. Both the quantities used to quantify the experience in either half have been scaled to a range of 0 to 10.

The regression summary is provided in the following table-

Variable	Estimate	Std. Error	t-statistic	p-value
Intercept	-1.4001	0.0679	-20.61	0
ϕ^{-1}(First Half Win %)	0.1255	0.0138	9.12	0
First Half Experience	0.206	0.0102	20.23	0
Second Half Experience	0.1776	0.0151	11.75	0

Table 5: Regression Model Summary:$\mathrm{R}^2 = 0.574$

4.4.3 Bootstrap Results

For performing resampling, 10000 bootstrap samples were drawn with replacement from the Lichess online database used for the study, each having size same as the original dataset. The regression was performed on all of the resampled datasets, and the bootstrap mean, variances and 95% confidence intervals for the regression coefficients are provided in the following table.

Variable	Estimate	Bootstrap Mean	Bootstrap Variance	Bootstrap C.I
Intercept	-1.4001	-1.4008	0.0034	(-1.566, -1.217)
ϕ^{-1}(First Half Win %)	0.1255	0.1255	0.0013	(0.0572, 0.1795)
First Half Experience	0.206	0.2061	0.0001	(0.193, 0.2183)
Second Half Experience	0.1776	0.1778	0.0003	(0.1324, 0.2009)

Table 6: Bootstrap regression summary for 10000 bootstrap estimates

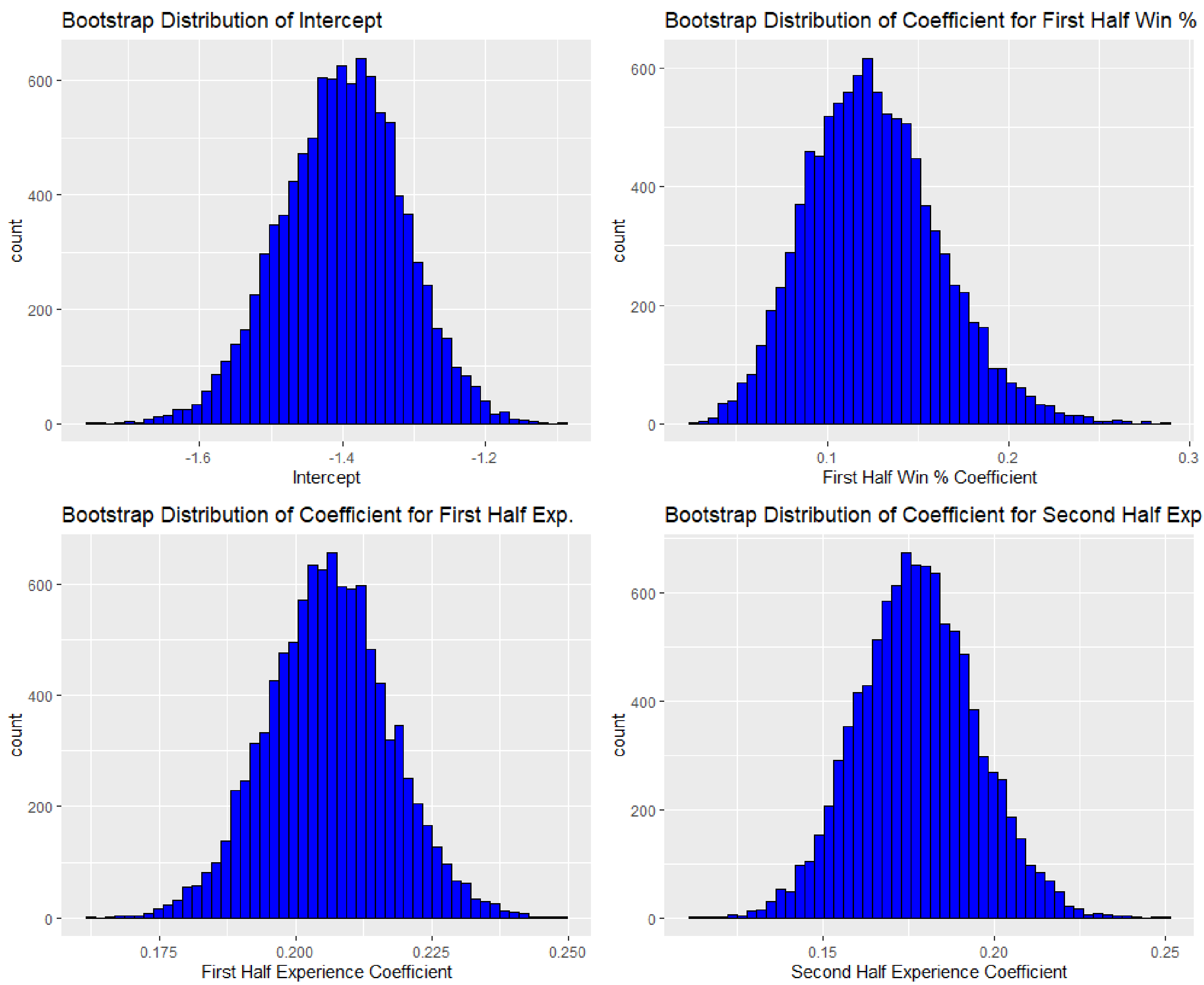

Figure 1: Histogram of Bootstrap Estimates of Regression Coefficients

5 Two player Rummy

For the analysis of this card game we use actual data from online games. We use data for 20,000 players, results are in table 7. The results show strong impact of experience in the face of performance. Current experience

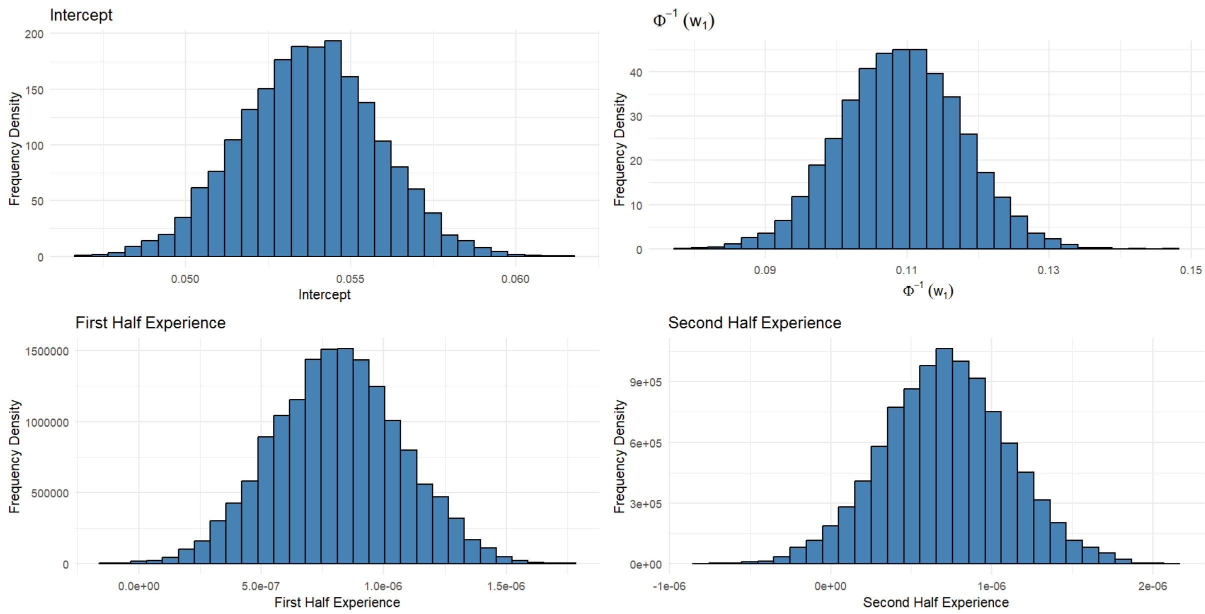

Figure 2: Bootstrapping Results for 2-Player Rummy Regression Model

also has some impact (t-statistic being more than 1).

Variable	Estimate	Std. Error	t statistic	p-value
Intercept	5.371e-02	1.938e-03	27.716	<2e-16 ***
$\phi^{-1}(w_1)$	1.092e-01	6.936e-03	15.738	<2e-16 ***
First Half Experience	8.109e-07	3.773e-07	2.149	0.0316 *
Second Half Experience	7.101e-07	6.669e-07	1.065	0.2870

Table 7: Regression Summary for Rummy, multiple R^2: 0.013, based on 20,000 sample users

Residual standard error: 0.2279, F-statistic: 88.33 on (3, 19996) df, p-value: 0.0000

5.1 Resampling results

Variable	Estimate	Bootstrap mean	Bootstrap variance	Bootstrap 95% CI
Intercept	0.05371	0.05371	4.4496e-06	$(0.0496, 0.0577)$
$\phi^{-1}(w_1)$	0.1091	0.10915	7.2609e-05	$(0.0924, 0.125)$
First Half Experience	8.109e-07	8.082e-07	0	-
Second Half Experience	7.101e-07	7.163e-07	0	-

Table 8: Bootstrap Regression Summary for $B = 10000$ bootstrap iterations for $n = 20000$ observations

6 Analysis for Ludo

6.1 Ludo Experimental Data

For Ludo, we have used pre-launch experimental data with a small number of players (see table 9. The regression results are presented in table 10. Here previous performance plays an important role, in addition we also see the number of games in the second tranche being a serious factor with a negative sign. This is probably due to fatigue, which we will investigate in more detail elsewhere. The positive contribution of previous experience is also mildly important.

Player ID	**1st Half Win %**	**2nd Half Win %**	**ϕ^{-1}(1st Half Win %)**	**ϕ^{-1}(2nd Half Win %)**	**1st Half Games**	**2nd Half Games**
2113	0.46	0.60	-0.09	0.25	26	25
2708	0.35	0.25	-0.38	-0.67	40	40
3179	0.52	0.62	0.04	0.31	68	67
3421	0.52	0.61	0.05	0.29	23	22
3544	0.31	0.37	-0.51	-0.32	41	40
3671	0.61	0.61	0.27	0.29	70	70
3276	0.50	0.60	0.00	0.25	40	40
3296	0.52	0.53	0.06	0.08	75	76
3449	0.53	0.55	0.08	0.12	72	73
3480	0.42	0.50	-0.19	0.00	26	26
3581	0.27	0.27	-0.60	-0.60	22	22
3646	0.45	0.49	-0.13	-0.03	41	41

Table 9: Data for Regression: First 6 IDS are from 24 moves, last 6 from 30 moves.

Coefficients	**Estimate**	**Std. Error**	**Test Statistic**	**p-value**
Intercept	0.34	0.14	2.39	0.04
ϕ^{-1}(First Half Win Proportion)	1.36	0.203	6.67	0.0001
First Half Games	0.16	0.086	1.84	0.103
Second Half Games	-0.16350	0.086	-1.89	0.095

Table 10: Linear Regression Summary:Multiple $R^2 = 0.878$

6.2 Ludo Online Data

This study uses data obtained from a new variant of the Ludo game as an integrated offering on the My11Circle ("MEC") platform, a fantasy gaming platform. The primary objective of launching the Ludo game on the MEC platform has been to collect the real user game play data. We use this data to conduct our Study. The dataset consists of a total of 265062 players, who have played a total of 3033645 games among themselves during 5^{th} June to 23^{rd} July. The subsetted dataset used for the actual analysis includes players who have played at least 10 games in both months, consists of a total of 16197 players, who have played a total of 2018935 games, and 1191656 games among themselves during 5^{th} June to 23^{rd} July.

Variable	Estimate	Std.Error	t-statistic	p-value
Intercept	-0.0988	0.0028	-35.038	0
ϕ^{-1}(First Half Win %)	0.2823	0.0081	34.866	0
First Half Experience	0.007	0.0048	1.459	0.143
Second Half Experience	0.0899	0.0047	18.928	0

Table 11: Regression Output: Multiple R^2 : 0.1364

The regression results firmly show that the role of previous experience (learning) and previous performance (innate skill) is quite prominent in determining current performance of a player. This is a strong indicator of the importance of skill in this Ludo game.

6.2.1 Bootstrap Results

To understand the robustness of our results, we now perform a 10000 fold bootstrap analysis of the regressions results.

Variable	Estimate	Bootstrap Mean	Bootstrap Variance	Bootstrap C.I
Intercept	-0.0988	-0.0985	0.00009	(-0.0991,-0.0982)
ϕ^{-1}(First Half Win %)	0.2823	0.2861	0.0013	(0.2819,0.2908)
First Half Experience	0.007	0.0062	0.00003	(0.0056,0.0072)
Second Half Experience	0.0899	0.0902	0.0001	(0.0893,0.0907)

Table 12: Bootstrap regression summary for 10000 bootstrap estimates

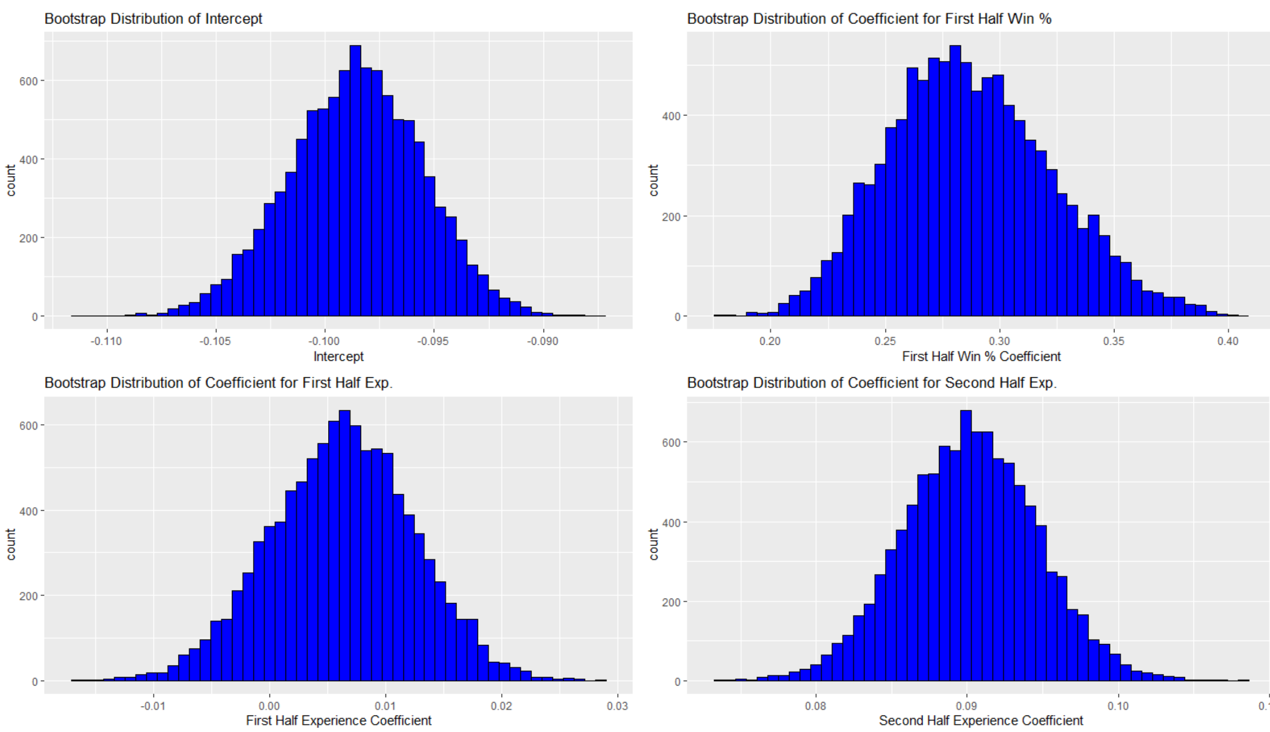

Figure 3: Histogram of Bootstrap Estimates of Regression Coefficients

The consistency of the results through the bootstrap analysis further strengthen our conclusions.

7 Teen Patti

We analyse the two most popular variations of this game, No Limit (result is in table 13) and Regular (result is in table 14). We also consider 5 player games only, as these are the most common. In both the versions, previous performance turns out to be an important factor which probably is also related to innate skills. The learning feature is supported by data as previous experience has a negative impact on current performance.

Variable	**Estimate**	**Std. Error**	**Test Statistic**	**p-value**
Intercept	-0.839	0.009	-89.827	0
ϕ^{-1}(First Half Win %)	0.357	0.004	81.99	0
First Half Experience	-5.716	0.589	-9.693	0
Second Half Experience	5.9159	0.595	9.936	0

Table 13: Regression Summary for **No Limit** games: multiple $R^2 = 0.116$ based on 53621 users

Variable	**Estimate**	**Std. Error**	**t-statistic**	**p-value**
Intercept	-0.922	0.0103	-89.515	0
ϕ^{-1}(First Half Win %)	0.226	0.006	38.077	0
First Half Experience	-0.886	0.723	-1.225	0.221
Second Half Experience	1.053	0.722	1.46	0.144

Table 14: Regression Summary for **Regular** Games: Multiple $R^2 = 0.038$ based on 38927 users

7.1 Resampling Analysis for Teen Patti

7.1.1 No Limit

Variable	Estimate	Bootstrap Mean	Bootstrap Variance	Bootstrap C.I
Intercept	-0.839	-0.838	0.0001	(-0.863,-0.814)
ϕ^{-1}(First Half Win %)	0.357	0.357	0.0001	(0.335,0.379)
First Half Experience	-5.716	-5.722	0.159	(-6.517,-4.949)
Second Half Experience	5.915	5.922	0.164	(5.142,6.728)

Table 15: Bootstrap regression summary for 10000 bootstrap estimates

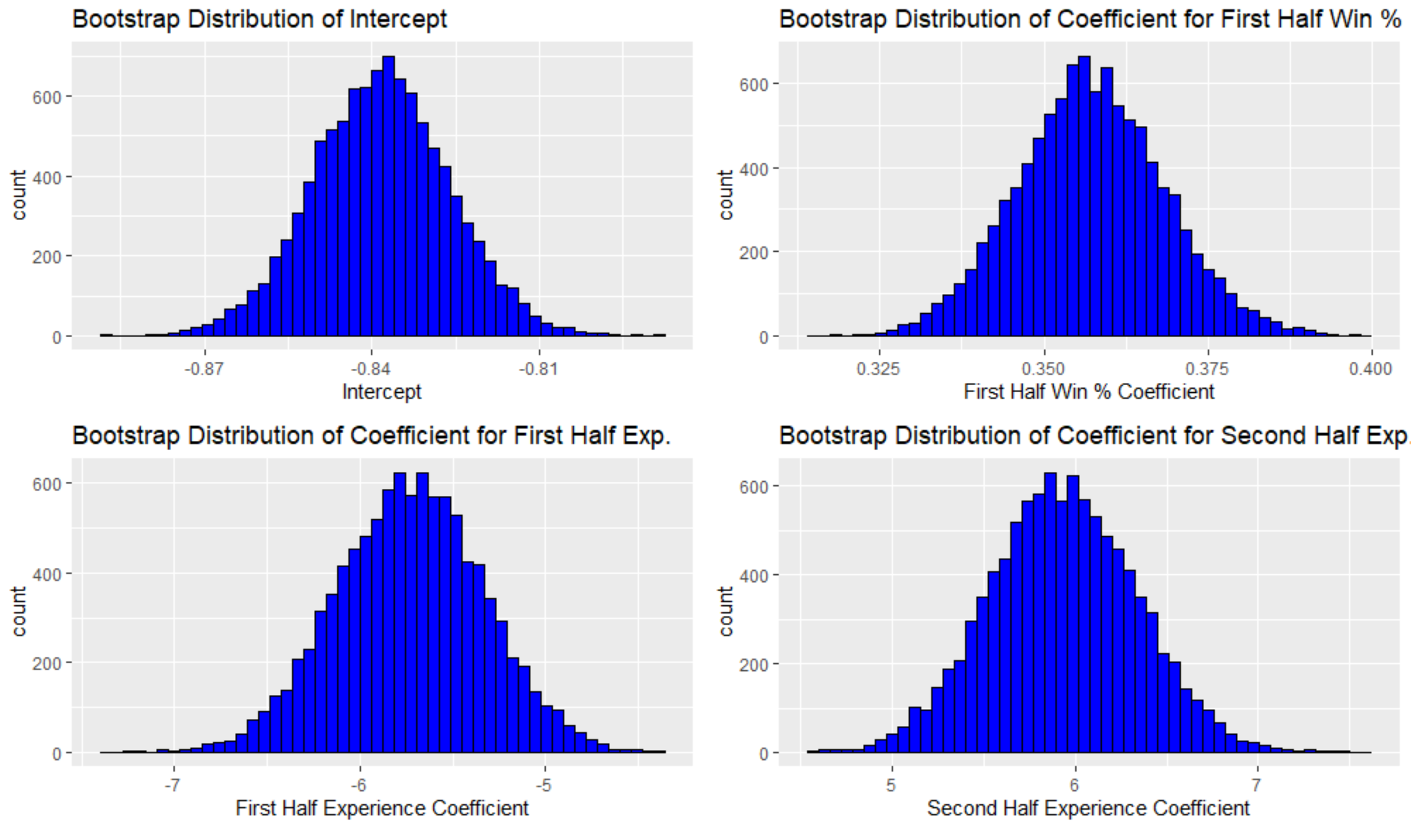

Figure 4: Histogram of Bootstrap Estimates of Regression Coefficients

7.1.2 Regular

Variable	Estimate	Bootstrap Mean	Bootstrap Variance	Bootstrap C.I
Intercept	-0.922	-0.922	0.0002	(-0.952, -0.891)
ϕ^{-1}(First Half Win %)	0.226	0.226	0.0002	(0.198, 0.255)
First Half Experience	-0.886	-0.898	0.544	(-2.196,0.505)
Second Half Experience	1.053	1.067	0.565	(-0.235,2.176)

Table 16: Bootstrap regression summary for 10000 bootstrap estimates

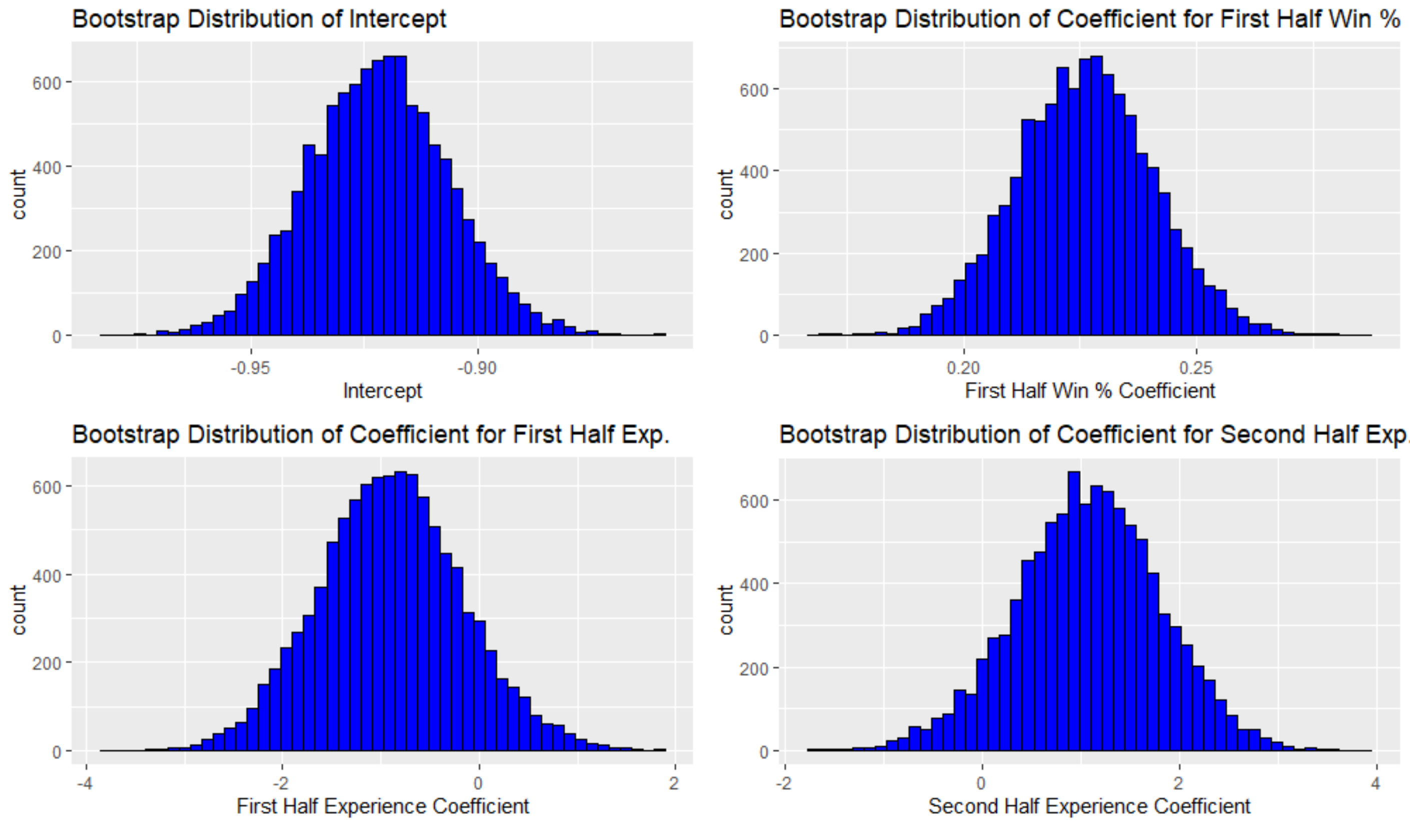

Figure 5: Histogram of Bootstrap Estimates of Regression Coefficients

8 Overall Discussion

8.1 Qualitative Considerations

In this report, our main aim is to draw comparative statements regarding the effect of Luck vs. Skill in the popular card and board games considered (tables 4 to 14). The initial discussion is around the importance of "innate skill" (performance in general) and "learning" (experience: learning by doing). Looking at the comparative results in table 17 we can assess this for the games under consideration. We certainly look at **Chess** as the benchmark case. Here rating proxies for skill acquired through experience. The regression results are very clear on the role of previous performance and rating.

Among the other three games, first consider **Teen Patti**, as this is popularly considered to be the "least skilled" game. Here we see a significant role of "innate skill" but no learning effect. This skill may be attitudinal (bluffing) or cognitive (recognising the behavioural markers of the opponent). But, there is no strategic learning.

Moving on to **Rummy** the results are very different. Again the relevant "innate skill" here is the choice of strategy in selection and retention of cards, and there is a significant learning effect. In fact this is true even for current experience (albeit at a lower level of significance). Contrasted with this, if we consider **Ludo**, we again see both skill as well as learning effect but the effect is now slightly weaker. In fact, the experience effect in Ludo is somewhat mixed with a positive and negative sign.

Game	Previous Performance	Previous Experience	Current Experience	R^2
Ludo				
Experiment	+	+ (*)	- (*)	0.88
Online	+	+ (*)	+	0.14
Chess (Real)				
Classical				0.48
Rapid (1y)	-	+	+ (Rating)	0.27
Rapid (2y)	+ (*)		+ (Rating)	0.61
Chess (Online)	+	+	+	0.57
Rummy (online)	+	+	+ (*)	0.013
Teen Patti (online)				
No Limit	+	-	+	0.15
Regular	+			0.05

Table 17: Overall Comparison (Significance at 5% level or *:$|t| > 1$)

We see that performance in all the games under consideration rely on innate skill (player heterogeneity), which is one indicator of skill base. Interim findings also suggest that an one shot game like **Teen Patti** is less amenable to learning compared to the repeated interaction games, e.g. **Rummy** and the two board games considered. Among the board games, **Ludo** illustrates a slightly weaker effect of learning compared to **Rummy** and, of course, **Chess** is the most clear cut skill based game. Based on our limited analysis, we can tentatively rank the four games in order of increasing skill and learning dependence as

$$\textbf{Teen Patti} \prec \textbf{Ludo} \prec \textbf{Rummy} \prec \textbf{Chess}.$$

8.2 The Skill Score

We now attempt a scoring exercise on a $[0, 1]$ scale for "Skill". This is tentative as it requires several subjective choices and can be confirmed only when we have more detailed game data on **Ludo** and **Chess**.

Again, we use the regressions results from tables 10 for Ludo, 4 for Chess in the Rapid format with 2 years data, 7 for Rummy and 13 for Teen Patti of the "No Limit" variety to represent the four games under consideration.

To avoid scaling issues, we use the t-statistics for the relevant factors instead of the regression coefficients as constituents of the score. Let

t_1 denote the t-statistics for performance in first half

t_2 denote the t-statistics for no. of games in first half

t_3 denote the t-statistics for no. of games in second half

Then, using the properties of the Normal distribution as we are working with fairly large samples, we define the normalised (in [0, 1]) versions as

$$x_i = \begin{cases} 0 & if \quad t_i < 0 \\ t_i/4 & if \quad 0 \le t_i \le 2 \\ (t_i+1)/6 & if \quad 2 < t_i \le 5 \\ 1 & if \quad 5 < t_i \end{cases}, i = 1, 2, 3.$$

The choice of the cut-offs, 2 and 5, are based on the level of significance (corresponding to 10^{-2} and 10^{-6} respectively). These are subjective choices and can be reconsidered. The choice of slope (1/4 and 1/6 is also subjective, the only objective decision being to award a higher importance to the initial phase (a concave evaluation).

Finally to define a skill score we propose that past experience should have twice the importance as innate skill, to give learning a higher weightage. In turn, innate skill in the form of past performance should also receive twice as much weight as current experience as current experience may be confounding learning and fatigue. Thus, we define the skill score as

$$s = \frac{2x_1 + 4x_2 + x_3}{7} \in [0, 1].$$

Of course, this formulation is just one possible among many and, end of the day, is a matter of subjective choice. We hope that we can objectify this choice with a more comprehensive analysis in future work.

Below we illustrate with the games considered here in table 18. The Scores obtained almost confirms our ranking posited above, the only exception being Ludo scores slightly higher than Rummy now. As these two numbers are very close to each other, we think that the results may be sensitive to the choice of cut-offs and weightages in the score formula.

Ludo		Chess		Rummy		Teen Patti	
t-stat (t)	scaled (x)	t-stat (t)	scaled (x)	t-stat (t)	scaled (x)	t-stat (t)	scaled (x)
34.866	1.00	9.12	1.00	15.738	1.00	93.692	1.00
1.459	0.37	20.23	1.00	2.149	0.52	-9	0.00
18.928	1.00	11.75	1.00	1.065	0.27	9	1.00
	$s = \mathbf{0.64}$		$s = \mathbf{1.00}$		$s = \mathbf{0.62}$		$s = \mathbf{0.43}$

Table 18: Skill Score for Games considered

8.3 Resampling Results on Skill Score

Based on the findings above, we have reason to investigate the choice of score function further. In this part of the report we consider various choices for the cut-off and weightage to make the functional form of the skill score flexible subject to the intuitive restrictions. To illustrate, we use a range of weightages (e.g. between 0.5 and 0.6 in place of 4/7).

One may also use different non-linear transformation $x_i(t_i) : R \to [0, 1]$ on the t-stat. As any smooth transformation on a suitable domain can be well approximated by piece-wise linear functions, we claim that by varying the cut-off and weightages, the proposed score function provides sufficient variety.

The results are not only dependent on the choice of the score function. Sampling fluctuation inherent in the data may also influence our scores and hence ranking. So we also carry out a bootstrap analysis with our regression model. We will get a range of t-stat based on the range of regression coefficients from bootstrap analysis. This will give us a range of t_i values and hence a range of x_i values. Thus, instead of a point estimate for the score function s for each game, it will have a range of values. We can then compare distribution of the score values to arrive at a comprehensive ranking template. Details follow.

Choice of cut-offs	Transformation $x_i(t_i) : R \to [0,1]$
2,5	$x_i = \begin{cases} 0 & if \quad t_i < 0 \\ t_i/4 & if \quad 0 \le t_i \le 2 \\ (t_i+1)/6 & if \quad 2 < t_i \le 5 \\ 1 & if \quad 5 < t_i \end{cases}, i = 1,2,3.$
1.5, 4.5	$x_i = \begin{cases} 0 & if \quad t_i < 0 \\ t_i/3 & if \quad 0 \le t_i \le 1.5 \\ (t_i+1.5)/6 & if \quad 1.5 < t_i \le 4.5 \\ 1 & if \quad 4.5 < t_i \end{cases}, i = 1,2,3.$
1.5, 5.5	$x_i = \begin{cases} 0 & if \quad t_i < 0 \\ t_i/3 & if \quad 0 \le t_i \le 1.5 \\ (t_i+2.5)/8 & if \quad 1.5 < t_i \le 5.5 \\ 1 & if \quad 5.5 < t_i \end{cases}, i = 1,2,3.$
2.5, 4.5	$x_i = \begin{cases} 0 & if \quad t_i < 0 \\ t_i/5 & if \quad 0 \le t_i \le 2.5 \\ (t_i-0.5)/4 & if \quad 2.5 < t_i \le 4.5 \\ 1 & if \quad 4.5 < t_i \end{cases}, i = 1,2,3.$
2.5, 5.5	$x_i = \begin{cases} 0 & if \quad t_i < 0 \\ t_i/5 & if \quad 0 \le t_i \le 2.5 \\ (t_i+0.5)/6 & if \quad 2.5 < t_i \le 5.5 \\ 1 & if \quad 5.5 < t_i \end{cases}, i = 1,2,3.$
(a,b), $a \in (1.5, 2.5)$ & $b \in (4.5, 5.5)$	$x_i = \begin{cases} 0 & if \quad t_i < 0 \\ t_i/2a & if \quad 0 \le t_i \le a \\ \frac{t_i+(b-2a)}{2(b-a)} & if \quad a < t_i \le b \\ 1 & if \quad b < t_i \end{cases}, i = 1,2,3.$

Table 19: Possible Transformations $x_i(t_i) : R \to [0,1]$

Note that the transformations posited in table 19 covers both convex and concave formulations. It is convex (concave) if $b > (<) 2a$. Thus, when we generalise the discussion to smooth functions, we can cover both the aspects. Further, we also allow a range of weightages in the table 20 which can approximate the domain under our intuitive restrictions quite well.

$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$
$\frac{2}{7}$	$\frac{4}{7}$	$\frac{1}{7}$
0.25	0.5	0.25
0.3	0.6	0.1
$\frac{a}{2}$	a, $a \in (0.5, 0.6)$	$1 - \frac{3a}{2}$

Table 20: Possible Range of Weightages

For each of the games we use 10000 bootstrap samples to generate the t-statistics. Hence, we get a distribution of scores for each of the combinations in table 19. These are presented in figures 6 for Chess, 7 for Rummy, 8 for Ludo and finally 9 for Teen Patti.

As expected, we now see a clearly articulated distribution of score for each game which shifts in centre and spread with the cut-offs and transformations chosen. Except for Chess, which is degenerate at 1. Now our skill quantification and comparison exercise transforms into a (possibly) partial ordering problem as the distributions for different games may now overlap.

To simplify our discussion, we produce the median and an 80% confidence interval for all the score distributions in table 21 where the degeneracy of Chess score is clear. It also clear that Teen Patti is less skilled than either Rummy or Ludo (the confidence intervals are practically separated) who are in turn less skilled than Chess. The comparison between Rummy and Ludo is unclear, as already guessed at earlier. The distributions overlap with Ludo having a little more spread and the centre of the distributions (median) also switching order between Rummy and Ludo for alternative parametric configurations. This clearly indicates that these two games are very similar in total skill content as indicated earlier. Of course, the kind of skill required may be quite diverse.

Weightage	(a, b)	Teen Patti	Rummy	Ludo	Chess
(0.25,0.5,0.25)	(2,5)	(0.3, 0.5, 0.5)	(0.496, 0.575, 0.636)	(0.500, 0.667, 0.826)	(1,1,1)
	(1.5,4.5)	(0.318, 0.5, 0.5)	(0.567, 0.639, 0.698)	(0.500, 0.722, 0.868)	(1,1,1)
	(1.5,5.5)	(0.318, 0.5, 0.5)	(0.561, 0.624, 0.671)	(0.500, 0.722, 0.838)	(1,1,1)
	(2.5,4.5)	(0.291, 0.5, 0.5)	(0.447, 0.518, 0.602)	(0.500, 0.633, 0.802)	(1,1,1)
	(2.5,5.5)	(0.291, 0.5, 0.5)	(0.447, 0.518, 0.582)	(0.500, 0.633, 0.784)	(1,1,1)
(0.26,0.53,0.2)	(2,5)	(0.308, 0.47, 0.47)	(0.505, 0.595, 0.659)	(0.470, 0.647, 0.816)	(1,1,1)
	(1.5,4.5)	(0.323, 0.47, 0.47)	(0.578, 0.657, 0.718)	(0.470, 0.706, 0.860)	(1,1,1)
	(1.5,5.5)	(0.323, 0.47, 0.47)	(0.576, 0.642, 0.689)	(0.470, 0.706, 0.829)	(1,1,1)
	(2.5,4.5)	(0.299, 0.47, 0.47)	(0.457, 0.536, 0.629)	(0.470, 0.611, 0.790)	(1,1,1)
	(2.5,5.5)	(0.299, 0.47, 0.47)	(0.457, 0.536, 0.606)	(0.470, 0.611, 0.772)	(1,1,1)
(0.28,0.57,0.15)	(2,5)	(0.317, 0.43, 0.43)	(0.517, 0.621, 0.694)	(0.430, 0.620, 0.802)	(1,1,1)
	(1.5,4.5)	(0.328, 0.43, 0.43)	(0.59, 0.68, 0.75)	(0.430, 0.683, 0.849)	(1,1,1)
	(1.5,5.5)	(0.328, 0.43, 0.43)	(0.589, 0.665, 0.683)	(0.430, 0.683, 0.816)	(1,1,1)
	(2.5,4.5)	(0.31, 0.43, 0.43)	(0.471, 0.56, 0.666)	(0.430, 0.582, 0.774)	(1,1,1)
	(2.5,5.5)	(0.31, 0.43, 0.43)	(0.471, 0.56, 0.641)	(0.430, 0.582, 0.754)	(1,1,1)
(0.3,0.6,0.1)	(2,5)	(0.323, 0.4, 0.4)	(0.525, 0.641, 0.721)	(0.400, 0.600, 0.791)	(1,1,1)
	(1.5,4.5)	(0.33, 0.4, 0.4)	(0.597, 0.699, 0.776)	(0.400, 0.667, 0.841)	(1,1,1)
	(1.5,5.5)	(0.33, 0.4, 0.4)	(0.596, 0.715, 0.738)	(0.400, 0.667, 0.806)	(1,1,1)
	(2.5,4.5)	(0.318, 0.4, 0.4)	(0.48, 0.579, 0.694)	(0.400, 0.560, 0.762)	(1,1,1)
	(2.5,5.5)	(0.318, 0.4, 0.4)	(0.48, 0.579, 0.667)	(0.400, 0.560, 0.741)	(1,1,1)

Table 21: For the four Games considered: Quantiles (q_{10}, q_{50}, q_{90}) of Score for different weightage and (a, b) combinations

8.4 Concluding Remarks

For consistency of comparison, in the end, we have only used data from online real player games based on a large sample of players and games. The analysis of other data sets, real or experimental, are only shown for illustrative purpose.

The conclusions are quite nuanced with Chess being the most skilled (extreme) and Teen Patti on the other end (but not an extreme score, around 0.45). While Rummy and Ludo are in between, the ranking between them is ambiguous. Depending on cutoffs and weightage schemes, the score distribution is always overlapping for this two with a wider range for Ludo than Rummy (possibly due to the larger sample of Rummy data).

References

[1] Diganta Mukherjee and Subhamoy Maitra, Unveiling the potential and scope of the Online Skill Gaming Industry: Study with technology students and professionals, *Report submitted to E-Gaming Federation*, 2023.

[2] Jakob Erdmann, The Characterization of Chance and Skill in Games, in *The Characterization of Chance and Skill in Games*, 2011.

[3] Michael Orkin, Games of Chance and Games of Skill, *Chance*, American Statistical Association, 2023. Available at: https://chance.amstat.org/2021/11/games/.

[4] Robert C. Hannum and Anthony N. Cabot, Toward Legalization of Poker: The Skill vs. Chance Debate, *UNLV Gaming Research & Review Journal*, 13(1):1-20, 2009.

[5] Tom Stafford and Nemanja Vaci, Maximizing the Potential of Digital Games for Understanding Skill Acquisition, *Current Directions in Psychological Science*, 31(1):49-55, 2022. DOI: 10.1177/09637214211057841.

[6] Max Knappe, and Renske Nijssen ,Eva van Agt, Maria Diez Perez. Chance and Skill in Games, *Master Research Project*, 2021.

[7] Peter Borm and Ben van der Genugten, On a relative measure of skill for games with chance elements, *Top*, 9:91-114, 2001.

[8] Subhamoy Maitra, Diganta Mukherjee and Swagatam Das, Role of Skill in the Game of Online Rummy: A Statistical Analysis, *DOI:10.13140/RG.2.2.12826.11203*, March 2023.

[9] S. D. Levitt and T. J. Miles, The role of skill versus luck in poker: Evidence from the World Series of Poker, *J. Sports Economics*, 15:31-44, 2014. DOI: https://doi.org/10.1177/1527002512449471.

[10] Adriaan Kalwij and Kris De Jaegher, The age-performance relationship for a cognitive-intensive task: Empirical evidence from chess grandmasters, *Sports Economics Review*, Volume 2, June 2023, Article 100010.

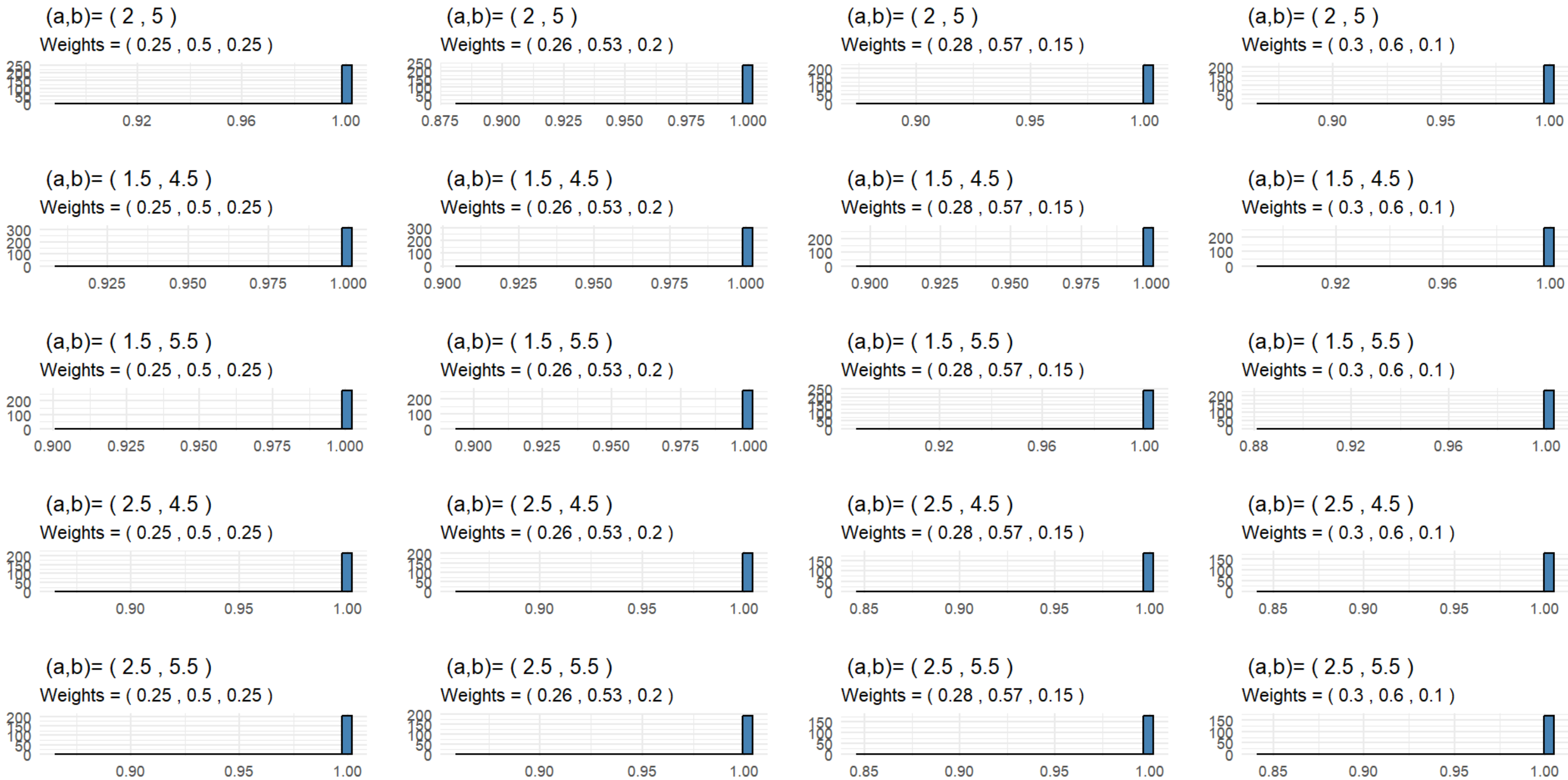

Figure 6: Distribution of Scores for Various Transformations for Chess Bootstrapped Data

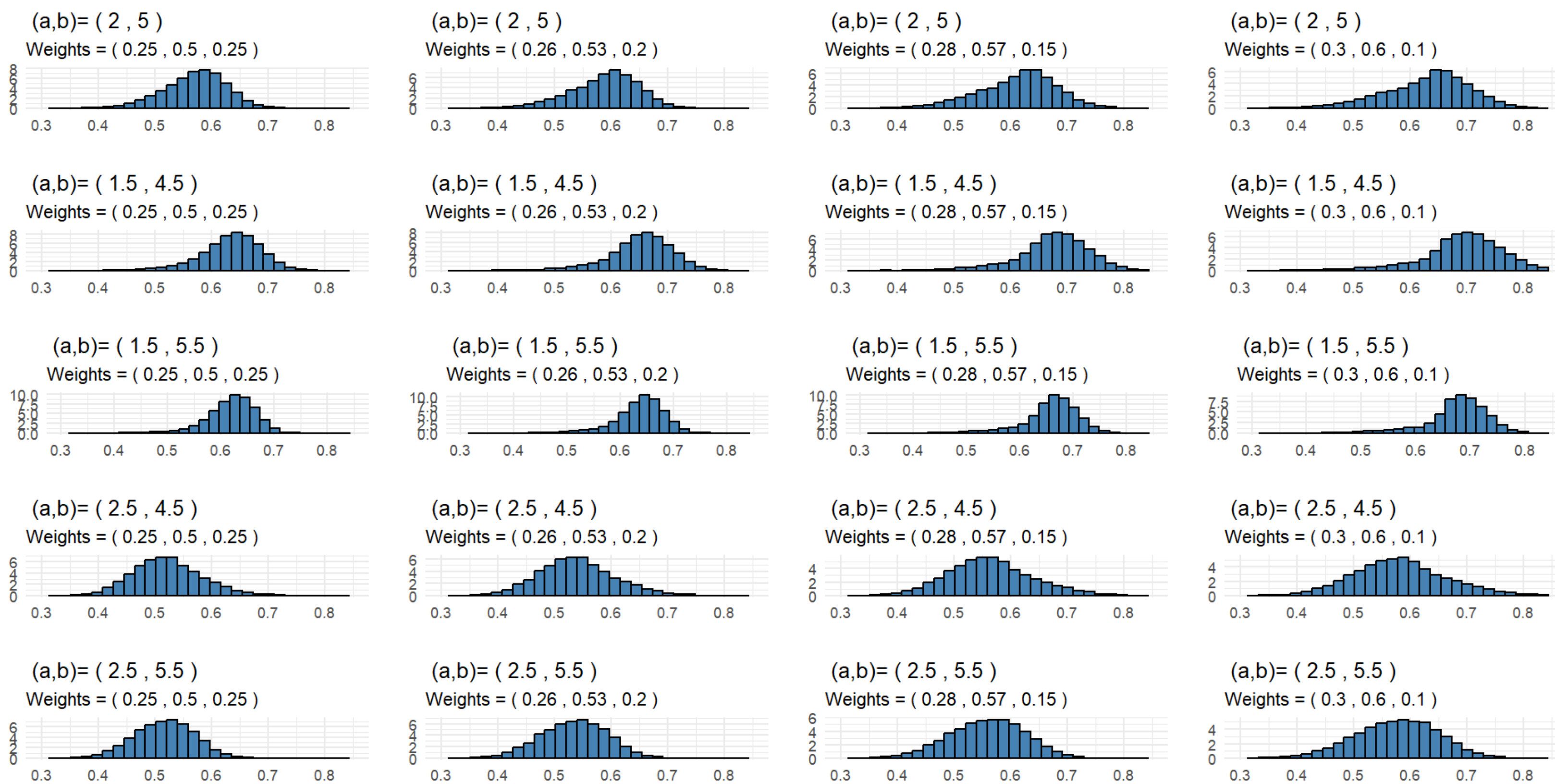

Figure 7: Distribution of Scores for Various Transformations for 2-Player Rummy Bootstrapped Data

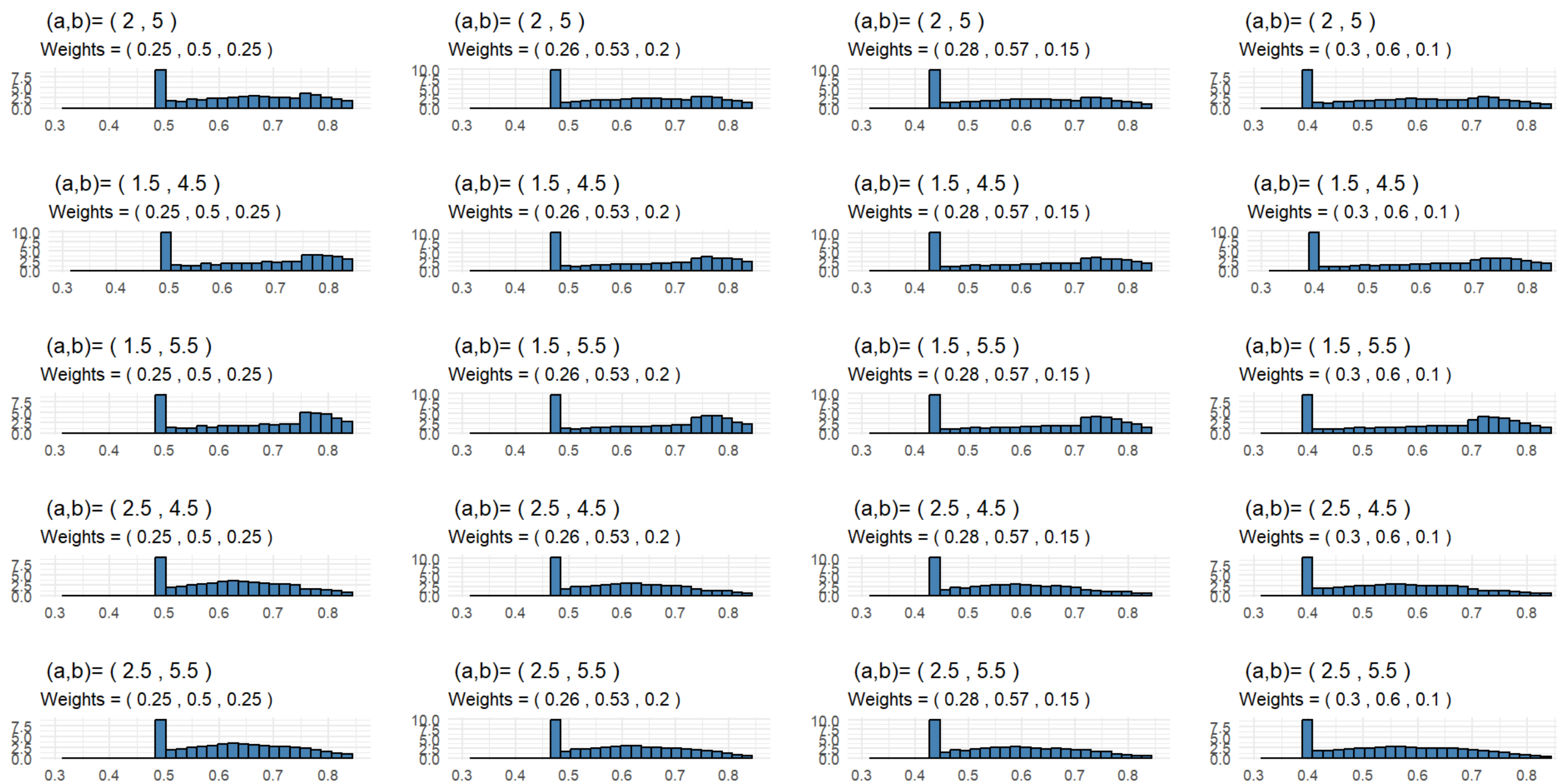

Figure 8: Distribution of Scores for Various Transformations for Ludo Bootstrapped Data

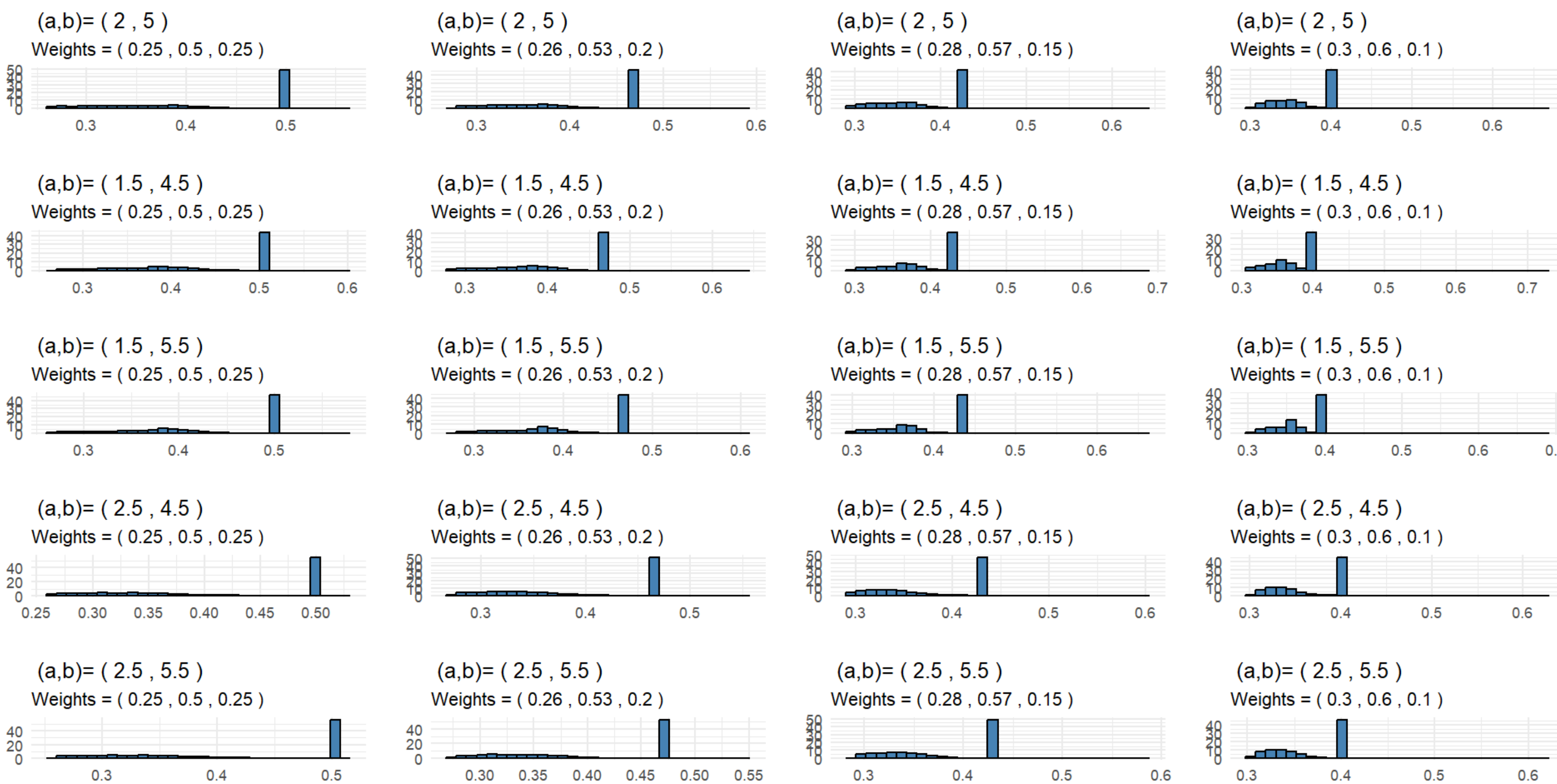

Figure 9: Distribution of Scores for Various Transformations for Teen Patti Bootstrapped Data