University of Heidelberg

Department of Economics

Discussion Paper Series | No. 687

Measuring skill and chance in different versions of Poker

Marco Lambrecht

August 2020

Measuring skill and chance in different versions of poker

Marco Lambrecht*

This Version: July 15, 2020

Abstract

This paper aims to measure skill and chance in different versions of online poker, using the best-fit Elo algorithm established in the first chapter. While *Texas Hold'em* arguably is the most popular version being played, the amount of skill involved might differ from other versions like *Omaha Hold'em*. Many platforms offer faster procedures to play (e.g. *"hyper turbo"*), as well as different levels of stakes. Given the richness of online poker data, it is possible to isolate the impact of these variations individually. The heterogeneity of best-fit Elo ratings decreases in quicker competitions or with higher stakes. Meanwhile, Omaha seems to contain more elements of skill than Texas Hold'em, as its analysis shows a wider distribution of skill levels of players.

Keywords: Elo-rating, measuring skill and chance, poker

JEL-Codes: L83, C72

*marco.lambrecht@awi.uni-heidelberg.de, University of Heidelberg, Department of Economics, Bergheimer Str. 58, D-69115 Heidelberg, Germany.

I like to thank Peter Duersch, Jörg Oechssler, Hannes Rau and Andis Sofianos for their helpful comments. Financial support by DFG grant OE 198/5-1 is gratefully acknowledged.

1 Introduction

Poker was originally played as a five card draw game, but over the years, other versions have become more popular, e.g. Texas Hold'Em, Omaha, Stud, and Razz (Fiedler and Wilcke, 2011). Casinos offer different stake levels and speeds of play. This work aims to compare the level of skill involved in these different versions of poker.

Previous researchers have worked on analyzing skill in poker. These approaches focus on Texas Hold'em (or simplified versions of it) and typically find it requires substantial skill. Dreef et al. (2003) theoretically define player and strategy types and compare their performance in simulations. DeDonno and Detterman (2008) instruct one group of subjects on how to play poker better and observe that this group outperforms the control group. An alternative approach is to compare poker to sports. Croson et al. (2008) compare data from poker to data from golf and find that past performances have similar predictive power in both games. My approach shows similarities to the studies of Fiedler and Rock (2009) and Potter van Loon et al. (2015), where comprehensive data of online poker is used to measure skill in poker. These studies find a significant amount of skill.

The approach in this paper is the closest to Duersch et al. (2020), who establish the best-fit Elo algorithm to empirically measure skill in games. Analyzing different datasets, the authors find that poker shows a small amount of skill compared to other games. Interestingly, they also find that the amount of skill in poker does not vary substantially with the number of players at the table.

To my knowledge, there have been no studies contrasting different versions of poker yet, as well as no studies that analyze variations in speed of play.

In this study, I implement the best-fit Elo algorithm (Duersch et al., 2020) to quantify the degree of skill and chance in different versions of two-player online poker. Specifically, I contrast *Texas Hold'Em* and *Omaha* poker, *regular* and *hyper turbo* settings, and stake levels of 3.50$, as well as 60$. The results show that heterogeneity of playing strengths in poker depend on the version played, speed, and

stakes. In general, faster play and higher stakes decrease heterogeneity, indicating a higher influence of chance. Additionally, Omaha exhibits a higher heterogeneity of skill than No Limit Hold'em.

The next section describes poker platforms, and the data used for the analysis. Section 3 provides the results, and section 4 concludes.

2 Data and Procedure

I analyze poker data that was purchased from a commercial vendor (HHSmithy) and was monitored on the online platform "PokerStars" between 2014 and 2017. The availability of data from commercial vendors exemplifies the popularity of different versions of poker. I focus on *Texas Hold'Em* and *Omaha Hold'em*. The Texas version features two private cards, while in the Omaha version players receive four private cards. In both versions, five community cards are laid out sequentially. These community cards are visible to every player and are common as every player can use them to form a poker hand. While in the Texas version, players can look for the strongest combination of five cards out of their private and community cards, in the Omaha version they have to use exactly two of their private cards and three of the community cards.

Additionally, PokerStars offers different speeds of play by varying chip endowments in tournaments. While *regular* (REG) tournaments start with 1500 chips per player, *hyper turbo* (HT) tournaments endow players with 500 starting chips. As the blind structure (i.e., the enforced bets in each hand) of both tournaments is the same[1], the smaller amount of starting chips leads to less wiggle room and thus (on average) to fewer hands needed to determine the winner of the tournament. Note that HT tournaments also impose a stricter time limit on each decision.

Clearly, poker can be played with different stakes. To investigate the impact on heterogeneity of playing strength, data from both medium stake (MS) 3.50$ and

[1]At the start of the tournament, the small blind is equal to 10 chips, and the big blind to 20 chips. Subsequently, blinds are increased at fixed intervals.

high stake (HS) 60$ tournaments are included in the analysis.

Overall, this study analyzes five datasets which are summarized in Table 1.

	Texas	Omaha
REG	MS	MS
HT	MS/HS	MS

Table 1: Poker data included in this study

The datasets consist of so-called "Heads-Up Sit-and-Go"-tournaments, which will start whenever two players sit down at the same table. Note that at this point both players pay the entry fee of the tournament. Then, they are endowed with an equal amount of chips and play until one player lost all his chips to the other player. Subsequently, the winner (i.e., the player holding all the chips) will be rewarded with money worth twice his entry fee.[2]

Table 2 summarizes the size of the different datasets. Specifically, it describes the datasets with respect to "Regulars", which are defined as players who have played at least 25 matches.

	#Matches	#Players	#Regulars	Max Matches (Regulars)	Mean Matches (Regulars)	Median Matches (Regulars)
Texas-REG-MS	191704	55158	1883	7531	126.2	49
Texas-HT-MS	370470	41540	5117	7961	113.7	55
Texas-HT-HS	69204	7661	617	2642	175.9	49
Omaha-REG-MS	13514	4804	153	1719	92.1	41
Omaha-HT-MS	256938	24188	2950	5285	147.5	61

Table 2: Statistics on matches, players and regulars in the different poker datasets

In order to analyze the data, this work applies the best-fit Elo algorithm as

[2]In fact, the casino will deduct a small amount of the prize money, the so-called "rake".

established by Duersch et al. (2020). The procedure involves the calibration of the Elo-rating for each dataset separately and rating each and every player accordingly.

The Elo-rating approximates playing strengths by assigning a rating to each player. Whenever two players meet in a competition at time t, their current ratings can be used to calculate expected winning probabilities,

$$E_{ij}^t := \frac{1}{1 + 10^{-\frac{R_i^t - R_j^t}{400}}}.$$

The rating R_i^t of player i is an empirical measure of player i's playing strength. More specifically, player i's chance of winning against j is dependent on the difference in ratings via the expected score $E_{ij}^t \in (0, 1)$, which can also be thought of as i's expected payoff (e.g. when a draw is counted as $\frac{1}{2}$).

The Elo ratings of the players i, j who are in match t are updated as follows,

$$R_i^{t+1} = R_i^t + k \cdot (S_{ij}^t - E_{ij}^t).$$

Here, S_{ij}^t denotes the observed score of player i in match t.[3] The ratings of players who are not involved in match t do not change. The best-fit Elo algorithm calibrates the parameter k for each dataset individually. In order to achieve the best possible calibration, the optimal value k^* is chosen as:

$$k^* := \arg\min_k \frac{1}{T} \sum_{t \in T} \left(S_i^t - E_i^t(k)\right)^2$$

Note that every tournament t results in two error terms, one for each player competing in tournament t. Intuitively speaking, k^* is chosen so that prediction errors (ex post) are minimized.

Once all players are rated according to the best-fit Elo algorithm, the focus is on the standard deviation of the rating distributions of each game. In the Elo rating, a given difference in ratings of two players corresponds directly to the winning probabilities when the two players are matched against each other. Thus, the more heterogeneous the ratings are, the better one can predict the winner of a match.

[3]Note that the Elo-rating is designed for situations where $S_{ij}^t \in [0, 1]$ and $S_{ij}^t + S_{ji}^t = 1$.

If the distribution of Elo ratings is very narrow, then even the best players are not predicted to have a winning probability much higher than 50%. The wider the distribution, the more likely are highly ranked players to win when playing against lowly ranked players, and the more heterogeneous are the player strengths. In my data, the rating distributions of all games are unimodal,[4] which makes it possible to interpret the standard deviation of ratings as a measure of skill. For further details on the best-fit Elo algorithm (including an extention to multiplayer cases), see Duersch et al. (2020).

3 Results

Following Duersch et al. (2020), I focus the analysis on "Regulars" (players that have played at least 25 matches).[5] Table 3 reports the result of the analysis. Note that the focus of measurement is the standard deviation of Elo rating distributions of regular players. The table reports the minimum and maximum rating, and the rating of the 1% and the 99% percentile player. One can transform the standard deviation of each game into the corresponding winning probability of a player who is exactly one standard deviation better than his opponent. This probability is denoted as p^{sd}. For comparison, the table also provides the winning probablities when a 99% percentile player is matched against a 1% percentile player, which is denoted as p_1^{99}. The winning probability p_1^{99} can be used to calculate the number of matches necessary so that a player who is in the top percentile wins more than half of the matches with a probability larger than 75% against an opponent that is in the bottom percentile.[6] This number is reported in the repetitions column

[4]See Figure 2 in the Appendix.

[5]Due to the updating nature of Elo ratings, initial ratings might not approximate true playing strengths well. On the other hand, setting the threshold too high might thin out the data. Figure 1 in the Appendix depicts the standard deviation of ratings for different thresholds of minimum games, where a threshold of 25 matches seems reasonable.

[6]This definition is used by Potter van Loon et al. (2015).

(abbreviated "Rep.").

	Std Dev	Min	1%	99%	Max	p_{sd}	p_1^{99}	Rep.
Texas-REG-MS	22.9	-98.6	-40.9	81.1	123.1	53.3	66.9	5
Texas-HT-MS	5.4	-28.5	-11.2	19.6	56.4	50.8	54.4	59
Texas-HT-HS	4.8	-13.6	-8.7	20.2	25.6	50.7	54.1	67
Omaha-REG-MS	28.6	-113.4	-75.6	61.6	64.6	54.1	68.8	3
Omaha-HT-MS	6.7	-60.7	-15.3	22.4	36.7	51.0	55.4	39

Table 3: Summary statistics on the distribution of Elo ratings in the different poker datasets. Note that, in contrast to chess, ratings are centered on zero by design.

Overall, the heterogeneity of playing strengths in the different versions of poker is moderate. This data suggests that even the best players do not win much more than two thirds of their matches when playing against the worst players.

Comparing the different versions of poker, it turns out that Omaha-REG-MS ranks in front of Texas-REG-MS. The same relative order of Omaha and Texas Hold'em persists when comparing HT-MS tournaments, implying that Omaha involves more skill than Texas Hold'em.

REG tournaments seem to involve significantly more skill than HT tournaments. Switching from Texas-REG-MS to Texas-HT-MS tournaments decreases the winning probability p_1^{99} from 66.9% to 54.4%. Similarly, comparing Omaha-REG-MS to Omaha-HT-MS, one can observe that p_1^{99} decreases from 68.8% to 55.4%.

Regarding different stake levels, Texas-HT-MS has a slightly wider distribution of ratings than Texas-HT-HS. While the differences in standard deviation and winning probabilities might not seem overly remarkable, one can see a (small) difference in repetitions needed for the better player to be ahead, changing from 59 to 67.

4 Conclusion

This paper investigates how the degree of skill differs in different versions of poker. Applying the best-fit Elo algorithm, which is designed to measure heterogeneity of skill empirically and comparable across games, I find that the amount of skill in different versions of poker clearly varies.

Comparing Texas Hold'em and Omaha, it turns out that Omaha contains more skill elements than Texas Hold'em. Comparing the rules of the Omaha and the Texas version, it is worth noting that Omaha players have additional private information due to the fact that they hold four private cards instead of two. This seems to increase the complexity significantly, for example when considering the calculation of winning probabilities before all community cards are revealed. Even after all cards are revealed, Texas Hold'em players have to evaluate 21 different five card combinations out of the community cards and their private cards to determine their strongest hand, while Omaha players have to consider 60 different combinations.[7]

Similarly, a variation in the speed of play influences the measured degree of skill in the expected direction. When reducing starting chips and putting time pressure on decisions, many tournaments are decided after only a few hands, often including all-in situations before any community card is revealed. This, by design, increases the influence of chance on the outcome.

Furthermore, my results show that an increase in stakes reduces heterogeneity. Note that, due to the fact that skill is measured in standard deviations of ratings, the fact that the player pool is smaller can not explain this finding. It might be conceivable though that players who feel confident enough to play for higher stakes have more experience, and thus (due to learning, see Fudenberg and Levine (1998)) might be closer to optimal play than their counterparts on medium stakes. Consequently, this selection effect would reduce the variation of skill levels within

[7]Compare combinations of five out of seven cards in Texas Hold'em, i.e. $\binom{7}{5}$, to combinations of two private and three community cards, $\binom{4}{2}\cdot\binom{5}{3}$, in Omaha.

this group.

Given the differences in the amount of skill involved in different versions of poker, one might wonder whether theoretical approaches that try to measure the amount of skill should (in an ideal world) try to account for some of the dimensions considered here. The differences in skill involved in REG and HT tournaments are quite remarkable, while it remains unclear how much of the difference can be attributed to the reduction of starting chips and how much to increased time pressure on decisions. Previous studies have shown that time pressure can lead to inefficient decision making (Zakay and Wooler (1984)), which might lead to smaller heterogeneity in skill while playing.

Due to the differences in popularity, the size of the datasets for the different versions is not balanced. Specifically, Omaha-REG seems to be played by few players, while in general HT tournaments are a lot more popular than REG tournaments. However, as Duersch et al. (2020) show, the best-fit Elo algorithm measures skill independent of the size of the dataset, conditional on a minimum amount of data to approximate well. One might argue that the smaller datasets of this study could be in question to meet that requirement.[8] It seems fair to assume though that the overall trend, namely that Omaha exhibits a larger amount of skill than Texas Hold'em, would not be reversed if more data was available.

In the context of professional poker players, it is worth noting that the profitability of playing a certain type of poker (while assuming to be a "winning player") not only depends on the winning probability, but also on the duration of play.

[8] I additionally analysed data for other poker versions, for example, *Fixed Limit Texas Hold'em*, but less than 40 players had played more than 25 matches. The high estimated standard deviation of 55.8 for this game version could offer support for a conjecture that Fixed Limit Texas Hold'em shows a larger heterogeneity in playing strengths than No Limit Hold'em. This can be because tournaments tend to be much longer and thus involve more decisions, since maximum bets are limited. According to Bowling et al. (2015), two-player Fixed Limit Texas Hold'em is solved from a game-theoretic perspective, which implies that players could potentially be playing almost optimally. Meanwhile, a game-theoretic solution for No-Limit Texas Hold'em seems beyond reach, but Moravčík et al. (2017) show that artificial intelligence can outperform human players.

Therefore, it is not clear whether Omaha players can potentially earn the most, as tournaments might take longer. The same argument might decrease the differences in profitability when comparing HT to REG tournaments, as one can potentially play several HT tournaments in the period of time it takes to play one REG tournament. Furthermore, variations in popularity of different versions influence the search costs to find a suitable tournament to play. In fact, given the relatively small amount of data that was available on Omaha-REG which reflects the frequency of play on the website, frictions might arise because of a limited number of opponents (and thus, opportunities to play).

5 Appendix

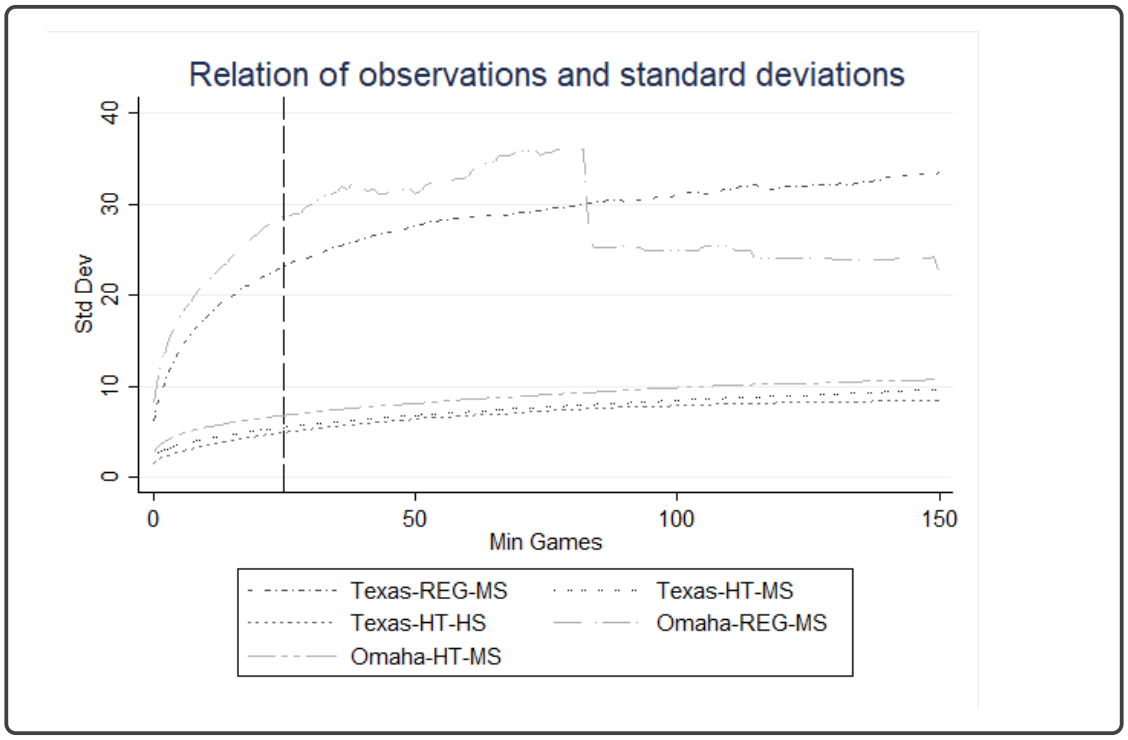

Figure 1: Standard deviation of rating distributions for different cut-off values (min. number of matches per player). The vertical dashed line indicates a minimum of 25 games.

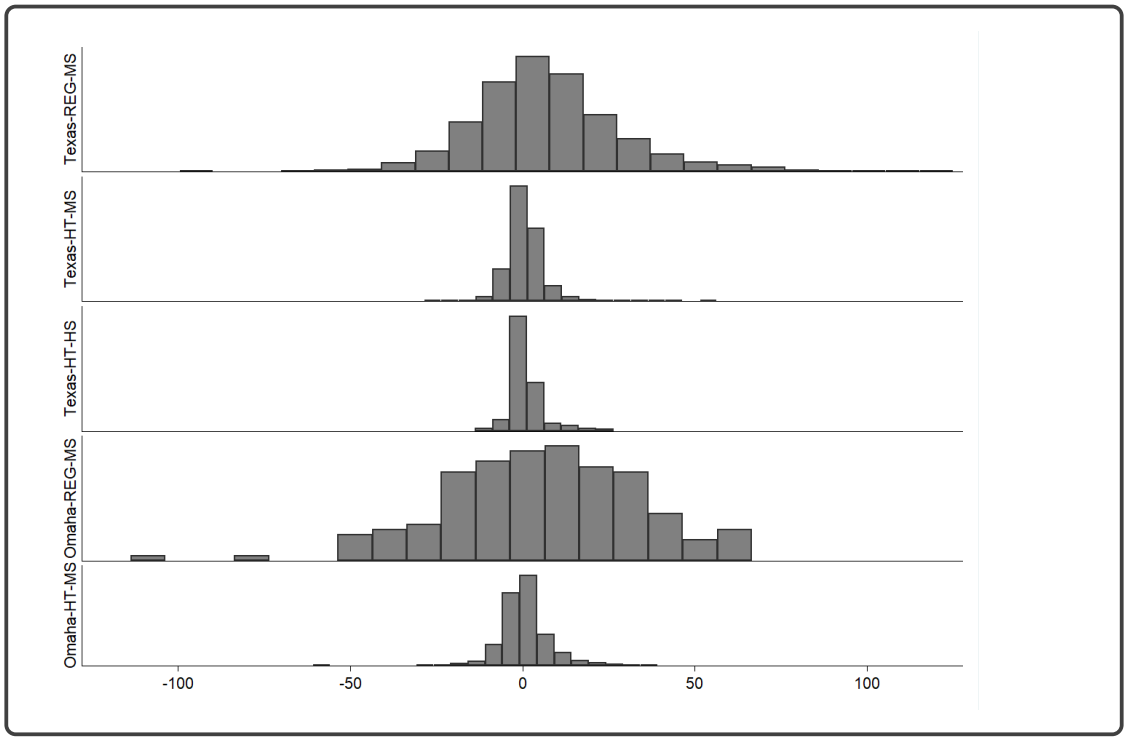

Figure 2: Rating distributions for different versions of poker.

References

Bowling, M., N. Burch, M. Johanson, and O. Tammelin (2015): "Heads-up limit hold'em poker is solved," *Science*, 347, 145–149.

Croson, R., P. Fishman, and D. G. Pope (2008): "Poker superstars: Skill or luck? Similarities between golf - thought to be a game of skill - and poker," *Chance*, 21, 25–28.

DeDonno, M. A. and D. K. Detterman (2008): "Poker is a skill," *Gaming Law Review*, 12, 31–36.

Dreef, M., P. Borm, and B. van der Genugten (2003): "On strategy and relative skill in poker," *International Game Theory Review*, 5, 83–103.

Duersch, P., M. Lambrecht, and J. Oechssler (2020): "Measuring skill and chance in games," *European Economic Review*, 103472.

Fiedler, I. and A.-C. Wilcke (2011): "The market for online poker," *Available at SSRN 1747646*.

Fiedler, I. C. and J.-P. Rock (2009): "Quantifying skill in games - theory and empirical evidence for poker," *Gaming Law Review and Economics*, 13, 50–57.

Fudenberg, D. and D. K. Levine (1998): *The theory of learning in games*, vol. 2, MIT press.

Moravčík, M., M. Schmid, N. Burch, V. Lisỳ, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, and M. Bowling (2017): "Deepstack: Expert-level artificial intelligence in heads-up no-limit poker," *Science*, 356, 508–513.

Potter van Loon, R. J. D., M. J. van den Assem, and D. van Dolder (2015): "Beyond chance? The persistence of performance in online poker," *PLoS one*, 10, e0115479.

Zakay, D. and S. Wooler (1984): "Time pressure, training and decision effectiveness," *Ergonomics*, 27, 273–284.