University of Heidelberg

Department of Economics

Discussion Paper Series | No. 643

Measuring Skill and Chance in Games

Peter Duersch, Marco Lambrecht and Joerg Oechssler

December 2017

Measuring skill and chance in games

Peter Duersch*
Department of Economics
University of Heidelberg

Marco Lambrecht†
Department of Economics
University of Heidelberg

Joerg Oechssler‡
Department of Economics
University of Heidelberg

This Version: December 20, 2017

Abstract

Online and offline gaming has become a multi-billion dollar industry. However, games of chance are prohibited or tightly regulated in many jurisdictions. Thus, the question whether a game predominantly depends on skill or chance has important legal and regulatory implications. In this paper, we suggest a new empirical criterion for distinguishing games of skill from games of chance: All players are ranked according to a "best-fit" Elo algorithm. The wider the distribution of player ratings are in a game, the more important is the role of skill. Most importantly, we provide a new benchmark ("50%-chess") that allows to decide whether games predominantly (more than 50%) depend on chance, as this criterion is often used by courts. We apply the method to large datasets of various two-player games (e.g. chess, poker, backgammon, tetris). Our findings indicate that most popular online games, including poker, are below the threshold of 50% skill and thus depend predominantly on chance. In fact, poker contains about as much skill as chess when 3 out of 4 chess games are replaced by a coin flip.

Keywords: ELO, ranking, games of skill, games of chance, chess, poker

JEL-Codes: L83, C72

---

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

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

‡oechssler@uni-hd.de, University of Heidelberg, Department of Economics, Bergheimer Str. 58, D-69115 Heidelberg, Germany.

We like to thank presentation audiences at Heidelberg, Dauthine Paris, ESA San Diego, GTM St Petersburg, HeiKaMaX, Free University Berlin, RTG Mannheim/Heidelberg and Fabian Krueger for their helpful contributions.

1 Introduction

Online and offline gaming has become a multi-billion dollar industry. According to the Economist, the legal gambling market amounted to more than 350 billion US dollar already in 2009 (Economist, 2010). The size of the industry justifies a careful investigation of the regulatory and economic issues that come with it.

From a legal perspective, a key aspect regarding this industry is what distinguishes games of skill from games of chance. This question has important legal and regulatory implications since in many jurisdictions games of chance are prohibited or tightly regulated, where one of the reasons given is the possibility of problem gambling and addiction. Furthermore, in most countries winnings from games are treated differently for tax purposes when they were generated in games of skill rather than in games of chance.[1]

So far, no convincing quantitative criterion exists that separates games of skill from games of chance. The difficulty arises because very few games are games of pure skill or games of pure chance. Mixed games, which involve both skill and chance elements, are by far the most popular games. Without much guidance from the theoretical literature, courts had to draw a line and often classify gambling as referring to games that "predominantly depend on chance".[2]

But how can one measure whether the outcome of a game depends predominantly on chance? Even if we all agree that predominantly means "more than 50 percent", the question is, "50 percent of what?"

Poker has been the most controversial game in this topic, especially because of its popularity. Hence, there has been an extensive debate in courtrooms as well as scientific journals as to whether poker is a game of chance or rather a game of

[1] For example, in the German tax code, see §4 Nr. 9b UstG

[2] 31 US Code §5362 targets "unlawful internet gambling"and defines betting and wagering in this context as "the purchase of a chance or opportunity to win a lottery or other prize (which opportunity to win is predominantly subject to chance)". Similarly, German law defines a game of chance subject to "the outcome depending largely or wholly on chance"(translated by the authors, §3 Abs. 1 GlueStV).

skill. Not surprisingly, different researchers (and courts) have come to very different conclusions. For example, in the US, several online poker providers were shut down in 2011 due to a violation of the Unlawful Internet Gambling Enforcement Act of 2006 (UIGEA).[3] In other jurisdictions like e.g. Austria, Israel, and Russia, poker is categorized as a game of skill (Kelly, Dhar, and Verbiest (2007)). In Germany, courts still refer to a decision by the Reichsgericht from 1906 that considered poker as a game of chance, while more recent courts considered the popular German card game "Skat" a game of skill.

In this paper we propose a new method for measuring the skill and chance components of games and apply it to poker, chess, backgammon, and several other popular games. The main objective of our measure is that it should provide a clear 50%-benchmark for the predominance of chance versus skill. Furthermore, it should be easily applicable to a variety of games and not be specific to one particular type of poker, say. Our approach is empirical and we benefit from the availability of very large data sets. Sport associations and online platforms track the outcomes of games played both online and offline. Thus millions of observations are available from public or commercial chess and poker data bases. We have also access to millions of observations from one of Europe's largest online gaming websites, which offers a variety of very different games, ranging from card games to crossword puzzles, from darts to football quizzes.

Our approach builds on previous research. It has long been argued and is now widely accepted that poker cannot be a game of *pure* chance. The basic idea is that in a game of pure chance (with time independent random devices like cards, dice, or roulette wheels) the past performance of players has no predictive power for their future performance. If past performance is found to have significant predictive power, this is a clear sign that skill does play a role for this game. There are several papers that take this approach and convincingly show that, for poker, skill plays a significant role (see e.g. Croson, Fishman, and Pope (2008), Levitt and Miles

[3]See United States Attorney, Southern District of New York (April 15, 2011).

(2014), and van Loon, van den Assem, and van Dolder (2015)).

Our approach differs from previous ones in two ways. First, rather than using performance measures like prize money won or finishing in the top x% in a tournament, we apply a complete rating system for all players in our data set. In particular, we use the Elo-system (Elo, 1978) used traditionally in chess and other competitions (e.g. Go, table tennis, scrabble, eSports). It has the advantage that players' ratings are adjusted not only depending on the outcome itself, but also on the strength of their opponents. Additionally, it is able to incorporate learning. The rating system can be applied to all games, even those that are not played for money. We calibrate the Elo rating system such that we get a best fit for each game. A given difference in ratings of two players has a direct correspondence in the winning probabilities when the two players are matched against each other. Thus, the more heterogeneous the ratings are, the better we can predict the winner of a match. The wider this distribution (measured by its standard deviation), the more heterogeneous are the player strengths. This is why we can interpret the standard deviation of ratings in a game as a measure of skill. Accordingly, the standard deviation is very high in games that are known to be pure skill and have a large heterogeneity of playing strength (e.g. chess). On the other hand, if the outcome of a game is entirely dependent on chance, in the long run, all players will exhibit the same strength of performance. Thus, the standard deviation of ratings tends to zero.

The second difference to the previous literature is that we propose an explicit 50%-benchmark for skill versus luck. We do this by constructing an artificial game that is arguably exactly half pure chance and half pure skill. For the pure skill part we use chess as a widely accepted a game of skill with the added benefit that there is an abundance of chess data. We construct our artificial game by randomly replacing 50% of matches in the chess data set by coin flips. This way, we mix chess with a game that is 100% chance and thereby construct what we call "50%-chess". We can now compare the standard deviations of ratings for all of our games to

50%-chess as a benchmark.[4]

Applying our method to the various datasets, we obtain a distribution of ratings for each game. As expected, chess has the highest standard deviation. Poker, on the other hand, has one of the narrowest distributions of all games. When we compare the games to our 50%-chess benchmark, we find that their standard deviations are mostly below the one from 50%-chess. Poker, backgammon, and other popular online games are below the threshold of 50% skill and thus depend predominantly on chance. In fact, when we reverse our procedure and ask how much chance we have to inject into chess to make the resulting distribution similar to that of poker, we find that poker contains about as much skill as chess when 3 out of 4 chess games are replaced by a coin flip.

There are a number of earlier approaches in the literature that mostly are concerned with poker. While most conclude that skill has a significant effect in poker, they do generally not quantify this effect. However, an interesting approach is to compare poker to sports or financial markets. Croson, Fishman, and Pope (2008) compare data from poker to data from golf and find that past performances have about the same predictive power in both games. Levitt and Miles (2014) calculate the return on investment of top players in the World Series of Poker and conclude that these are comparable to or even higher than returns in financial markets (concluding that either both are games of skill or none).

Several studies try to define certain player or strategy types and compare their performance in simulations or experiments. Borm and van der Genugten (2001), Dreef, Borm, and van der Genugten (2003, 2004a,b), and van der Genugten and Borm (2016) propose measures that compare the performances of different types of players. In order to calculate which part of the difference in performance may be attributed to skill and which to chance, they include as a benchmark an informed hypothetical player who knows exactly which cards will be drawn. The use of their

[4]One may argue that chess outcomes are still somewhat random and therefore it might not be the perfect reference point for pure skill. However, in this case our approach would still supply a lower bound for the 50% threshold.

approach is, however, limited to simplified versions of poker. Nevertheless, even for simple poker variants, the different studies report a substantial degree of skill.

Larkey, Kadane, Austin, and Zamir (1997) and Cabot and Hannum (2005) conduct simulation studies with different strategy types and find that more sophisticated strategies perform better. DeDonno and Detterman (2008) give one group of subjects some instruction on how to play better poker and observe that this group outperforms the control group. Siler (2010) shows that performance in online poker is related to playing style (aggressive, tight etc.), and that differences in style and performance between players decrease as stakes increase.

Finally, if a game has a skill component, in the long run by the law of large numbers better players will outperform weaker players. Thus, one way of measuring the skill component is to calculate how long it takes for a better player to be ahead of a weaker player with a certain probability. Fiedler and Rock (2009) propose a "critical repetition frequency" and find that it takes about 750 hands of online poker in their data for skill to dominate chance. Similarly, van Loon, van den Assem, and van Dolder (2015) use simulations to calculate the minimum number of hands for a player who ranks in the top 1% to outperform a player who ranks in the worst 1% with a probability $p > 0.75$. They find that the threshold is about 1500 hands. Our preferred measure can also to be expressed in terms of frequency of play and we report the according numbers below.

The rest of the paper is organized as follows. In section 2 we explain our new approach for measuring skill and chance in detail. Section 3 describes our data and in Section 4 we present the empirical results. Section 5 concludes.

2 A new approach for measuring skill and chance

The basic idea of our approach is not new and was used by a number of authors (see e.g. Croson, Fishman, and Pope (2008), Levitt and Miles (2014)). It is an empirical approach that involves checking whether the past performance of players can predict their future success. In a game of pure chance, the past has no predictive

power for the future (if the random draws are time independent). If a particular player was successful in roulette, this does not imply that they will be successful in the future. In a game of skill, this is obviously different. As our measure of past success, we use the ELO rating (Elo, 1978). It is well-established and can be applied to all games, even if no money is involved.[5]

Thus, the first step in our procedure is to rate all players in all games according to a simple Elo rating formula. This formula has one parameter that needs to be calibrated for each game. In subsection 2.1 we explain in detail how this is done. Once all players are rated, we can look at the distribution of player ratings for a given game. The wider this distribution (measured by its standard deviation), the more heterogeneous are the player strengths. Elo rating differences of players correspond to their predicted winning probabilities via a logistic function.[6] Therefore, the heterogeneity of ratings is correlated to the predictability of outcomes and is a proxy for the amount of skill involved. In a game of pure chance, the theoretical standard deviation of ratings is zero, as the past cannot predict the future.[7] In a game of pure skill like chess, the standard deviation is very high.[8]

The standard deviations of ratings give us an ordinal measure as it allows us to make statements like "game A is more of a skill game than game B". Our aim, however, is to define a general measure of skill and chance in games that

[5]Many other ways to measure past success are possible, see, e.g. Croson, Fishman, and Pope (2008) and Levitt and Miles (2014).

[6]Elo's original proposal (Elo, 1978) was based on a normal distribution. Today, the United States Chess Federation (USCF) uses a logistic function based on an extreme value distribution (see Glickman (1995)). The Fédération Internationale des Échecs (FIDE) still assumes normal distributions.

[7]Practically, it takes marginal positive values due to the skewness of stochastic outcomes, but it approaches zero while the number of observations increases.

[8]Note that even in chess the outcome is not perfectly deterministic (which would correspond to an infinite rating difference between any two matched players). In fact, deterministic outcomes, such as determining the winner via "who is older?", are no fun to play and are not commonly regarded as "games".

allows to specify whether a game is "predominantly" a game of skill or chance, respectively. For this purpose our innovation is to construct an artificial game that is a convex mixture of chess and a coin flip. Chess is commonly regarded as an archetypical game of skill. It is also widely known and very large data sets are available, making it a good benchmark. A coin flip, on the other hand, is an archetypical game of chance. We construct our artificial game "x%–chess" by replacing randomly $(100-x)$% of matches in our chess data by a coin flip. In fact, since chess has many draws, we allow our coin flip to have a "draw" as well. Thus, we replace the outcomes of the chosen matches by a "draw" with probability γ, where γ is the fraction of draws in the original chess data set, by a "win" with probability $\frac{1}{2}(1-\gamma)$ and a "loss" with probability $\frac{1}{2}(1-\gamma)$.

In most cases we will use "50%-chess" as our benchmark since this is the common interpretation of "predominantly skill" used by courts and legislators around the world.[9] Thus, if the standard deviation of a given game is higher than that of 50%-chess, we will say that the game is predominantly skill. If it is below, it is categorized as a game of predominantly chance.

2.1 Calibrating the Elo ratings

The Elo-rating (Elo, 1978) is defined for two-player games. As data we have a finite set of players I to be ranked, a finite number of matches T, and a finite series of outcomes from each match $t \in \{1, ..., T\}$ between players i and j, where $i, j \in I$. Outcomes are denoted by $S_{ij}^t \in [0, 1]$ and can, for example, be a win for player i ($S_{ij}^t = 1$), a loss ($S_{ij}^t = 0$), or, a draw ($S_{ij}^t = 0.5$). In some games intermediate outcomes may be allowed. Due to the constant-sum nature of the outcomes, it holds that $S_{ji}^t = 1 - S_{ij}^t$. We denote the set of players involved in match $t \in T$ by $\rho(t)$.[10]

The rating R_i^t of player i is an empirical measure of player i's playing strength.

[9] cf. footnote 2.

[10] In our case, this is always a pair of players.

More specifically, player i's chance of winning against j is related to 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 $\frac{1}{2}$) and is given by

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

Expected scores range from zero (sure loss) to one (sure win). The parameter 400 in the logit function is an arbitrary normalization used by chess federations which we retain for familiarity. Given this parameter, a rating difference of 100 translates into an expected score of .64.

We normalize the initial rating of each player to $R_i^0 = 0$.[11] The Elo ratings of the players who were involved in match t are updated as follows, [12]

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

$\forall i, j \in \rho(t), j \neq i$.

While the actual scores S_{ij}^t are observed in our data, the expected scores, are determined recursively and depend on k. To indicate this we use the long hand form $E_{ij}^t(k)$. A crucial element of the procedure is the determination of an appropriate value for k. This so–called k-factor determines by how much ratings are adjusted after observing a deviation of the actual score from the expected score in each match. Clearly, there is a trade-off between allowing for swift learning on the one hand and reducing fluctuations of rankings due to the inevitable randomness of outcomes in games with stochastic outcomes. In reality, the k-factor is chosen in many different, complicated, and relatively ad hoc ways by the different sports and chess federations.[13]

[11] Typically, chess federations use a positive initial rating. However, since only rating *differences* matter, this normalization is without loss of generality.

[12] The ratings of players who are not involved in match t do not change, $\forall i \notin \rho(t) : R_i^{t+1} = R_i^t$.

[13] For instance, the United States Chess Federation (USCF) historically used a set of fixed k-factors, where the value for each player was chosen according to his present rating. Today, they calculate the k-factor for each player separately depending on his rating in a quite complex way (for details, see Glickman and Doan (2017)).

Our approach is to calibrate the k-factor for each game in order to obtain the best fit given our data. The goal is to predict the winning probabilities as accurately as possible. For this purpose we minimize the following quadratic loss function summing over all matches of all players:[14]

$$k^* := \arg\min_k \frac{1}{T} \sum_{\substack{t \in T \\ i,j \in \rho(t)}} \left(S_{ij}^t - E_{ij}^t(k) \right)^2 . \tag{1}$$

We derive the solution to this minimization problem numerically.[15]

It may be tempting to interpret a high k^*-factor as a sign of a game of skill. A game of pure luck would produce a k-factor of essentially zero since past ratings have no predictive power for future winning probabilities. However, there are two reasons why the k-factor is an undesirable measure of skill. First, the learning curve can differ from game to game. In some games, learning will be slow and gradual. In other games, learning could be condensed into a single "epiphany" (Dufwenberg et al., 2010). The k-factor of these different types of games is likely to be very different although both may be games of skill. Second, the optimal k-factor depends on the number of observations in the data. This is so because of the above mentioned trade-off between swift learning and reducing fluctuations. Our preferred measure, the standard deviation of ratings, does not suffer from these drawbacks.

3 Data

In order to apply the proposed measure in practice, we acquired large datasets of various two-player games. These include competitions of chess, poker, and various online browser games. The size of the datasets as well the distribution of matches among players differ and are summarized by the statistics in Table 1. For each

[14]Note that each match produces two error terms in (1). However, given that $S_{ji}^t = 1 - S_{ij}$ and $E_{ji}^t(k) = 1 - E_{ij}^t(k)$, the solution to the minimization problem when considering only one error for each match is the same.

[15]See Appendix for an exact description of the numerical procedure.

	Chess	*Tetris*	*Jewels*	*Rummy*	*Solitaire*	*Backgammon*	*Yahtzee*	*Crazy eights*	*Poker*
#Players	235,110	10,872	39,058	7,851	33,860	4,279	10,079	12,557	58,806
#Regulars	18,963	139	1,899	108	3,297	179	444	302	446
#Matches	4,254,657	47,718	441,996	39,448	641,278	42,155	106,800	102,415	194,032
Mean Matches	36.2	8.8	22.6	10.0	37.9	19.7	21.2	16.3	6.6
Median Matches	11	2	5	3	6	4	4	4	1
99% Matches	411	111	281	125	448	225	305	160	79
Max Matches	2,280	514	1,649	3,026	3,277	1,301	1,378	2,945	7,531
Std. Dev.	87.7	23.0	22.6	45.5	105.8	54.9	68.7	50.5	65.1

Table 1: Statistics on players and their number of matches

game, it lists the number of players and the number of "regulars". The latter are those players who play at least 100 matches within our data. Furthermore, we report the total number of matches, the mean number of matches of each player, the number of matches of the median player as well as the 99% percentile player and the maximum number of matches played by a single player. Eventually, we list the standard deviation of the distribution of matches among all players.

Regarding chess, we were able to obtain a fairly comprehensive database provided by ChessBase. The observations date back to 1783 and include nearly 5 million matches in total. We restrict ourselves to a subset of the data ranging from 2000 to 2016, excluding any rapid and blitz formats.[16] The resulting subset consists of roughly 4.25 million matches from more than 235,000 players.

The poker data consist of so-called two-player "heads-up" *Sit-and-Go*-tournaments (SnG), a competition type where players pay an equal entry fee, are endowed with an equal stack of chips, and compete until all chips are owned by one player. We bought the data from "HH Smithy", a commercial provider of poker hand histories. The data we measure for this project include 58,806 players who participate in 194,032 tournaments. They took place between February 2015 and February 2017. All of these tournaments are "Texas Hold'em"competitions, which is the most popular type of poker online. The entry fee for each amounts to $3.50.

[16]These types of chess have more restrictive time limits for the players and are usually separated from "standard"chess, i.e. chess federations use separate ratings for these formats.

In addition, we acquired data from one of Europe's largest online gaming platforms, where a variety of games can be played in a web browser for money. The dataset includes more than 13 million matches in total, from more than 35 different games. We restrict the analysis to deal with games that are (more or less) well-known, or comparable to well-known games, giving us more than 1.25 million competitions. The number of different players for each game range from about 4,000 to 40,000. The games used are online versions of rummy, tetris, backgammon, jewels, solitaire, yahtzee and crazy eights.[17]

4 Results

In our result tables we report statistical values about the calculated ELO rating distributions. These include the minimum and maximum rating, the rating of the 1% and the 99% percentile player, and most importantly, the standard deviation of all ratings. We sort the tables according to this value. Furthermore, we 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. We refer to this probability as p^{sd}. For comparison, we also provide the winning probability of the 99% percentile player when competing with the average player, which we call p^{99}. The winning probability p^{sd} can be used to calculate the number of matches necessary so that the better player "most likely"is ahead. Formally, this means that a player who is one standard deviation better than his opponent wins more than half of the matches with a probability larger than 75%.[18] This number is reported in the repetitions column (abbreviated "Rep."). In Table 2, we also report the mean rating difference of players entering a match.

Table 2 provides the results when measuring all players in our database. Inspecting their rating distributions and comparing the standard deviations of the analyzed games, we find that most of them are substantially below the bench-

[17]For a detailed description of these games, see Appendix 6.1.

[18]This definition is used by e.g. van Loon, van den Assem, and van Dolder (2015)

	Min.	Max.	1%	99%	Std. Dev.	p^{sd}	p^{99}	Rep.	Mean Diff.
Who is older	-∞	∞			∞	100.0	100.0	1	∞
Chess	-684.6	945.3	-247.6	439.7	123.4	67.0	92.6	5	112.9
Tetris	-371.4	373.9	-120.4	180.8	52.4	57.5	73.9	21	72.3
50% Chess	-205.9	312.6	-58.0	109.8	28.1	54.0	65.3	71	36.0
Jewels	-411.1	225.0	-79.1	80.0	27.1	53.9	61.3	75	39.3
Rummy	-137.3	121.7	-37.4	55.5	14.7	52.1	57.9	259	37.0
Solitaire	-176.8	122.5	-40.0	48.2	14.1	52.0	56.9	285	22.2
Backgammon	-120.1	130.1	-32.8	42.4	12.3	51.8	56.1	351	25.6
Yahtzee	-65.3	86.3	-26.0	38.4	9.7	51.4	55.5	581	22.4
Crazy eights	-106.0	185.8	-18.2	23.7	7.5	51.1	53.4	941	12.8
Poker	-98.9	123.5	-12.0	19.9	6.0	50.9	52.9	1,405	30.2
Coinflip	0	0	0	0	0	50.0	50.0	∞	0

Table 2: Results all players

mark of 50%-chess. The online version of Tetris is the single game that exhibits a larger heterogeneity of skill and positions itself above the threshold. Moreover, the browsergame Jewels appears to be on a similar level as 50%-chess, as their standard deviations differ only slightly. Poker, on the other hand, stands at the bottom of the list. In terms of heterogeneity of skill, it seems to be inferior to the benchmark as well as many of the online games we analyzed. Regarding winning probabilities, a poker player who is one standard deviation better than his opponent seems to have a 50.9% chance of winning the competition. This translates into more than 1400 repetitions that are needed for the better player to most likely be ahead of his opponent.

For comparison, Table 3 shows our results when considering only the subset of ratings of "regulars". This group exclusively consists of players that have competed 100 or more times. When restricting the distributions to regulars, the order of games changes slightly. While passing several browser games (i.e. crazy eights,

	Min.	Max.	1%	99%	Std. Dev.	p^{sd}	p^{99}	Rep.
Who is older	-∞	∞			∞	100.0	100.0	1
Chess	-684.6	945.3	-200.9	703.4	188.3	74.7	98.3	3
Tetris	-225.2	306.3	-154.5	260.2	84.5	61.9	81.7	9
Jewels	-411.1	225.0	-205.5	128.8	58.2	58.3	67.7	17
50% Chess	-205.9	312.6	-74.1	203.1	56.4	58.0	76.3	19
Rummy	-137.3	121.7	-100.0	118.3	46.5	56.7	66.4	25
Backgammon	-120.1	130.1	-86.1	106.1	36.4	55.2	64.8	43
Poker	-98.9	123.5	-53.3	104.3	31.0	54.4	64.6	59
Solitaire	-176.8	122.5	-80.6	76.9	30.7	54.4	60.9	59
Yahtzee	-65.3	86.3	-55.5	77.8	29.2	54.2	61.0	65
Crazy eights	-106.0	185.8	-50.5	60.7	26.8	53.8	58.6	79
Coinflip	0	0	0	0	0	50.0	50.0	∞

Table 3: Results regulars

yahtzee and solitaire), poker remains in the lower part of the list. It is still clearly below the benchmark. In addition, the standard deviations of all games increase substantially. This is a consequence of the fact that our dataset includes many players who only compete in very few matches. Because of the updating characteristic of the Elo ratings, the deviation from the initial rating is limited subject to the number of matches a player competes in.[19]

The histograms in Figure 1 provide the full distributions of chess, 50%-chess and poker. Comparing the distributions of poker to those of chess and 50%-chess, it is apparent that the heterogeneity of ratings is much smaller for poker. The rating distributions of regulars show a right-shift compared to those of all players. Additionally, the increase in standard deviations is observable.

We can now also reverse our procedure and ask: how much chance do we have

[19]Besides that, the possible range of the final rating of a player also depends on the ratings of his opponents.

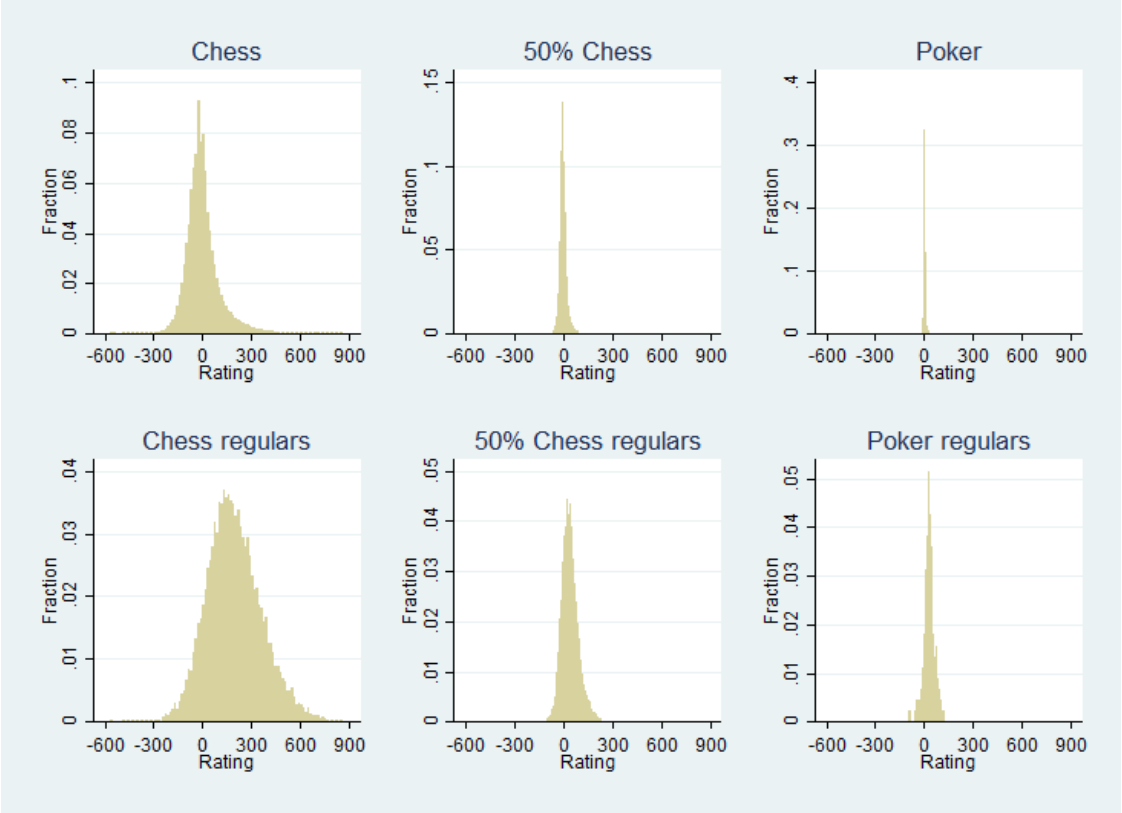

Figure 1: Rating distributions for chess, 50%-chess and poker for all players, as well as restriction to regulars

to inject into chess to obtain a distribution of player ratings similar to poker. As a result, we find that we would have to replace roughly 3 out of 4 chess games by a coin flip in order to produce a rating distribution as the one in poker.

Result 1: Of all the games we consider only 1 or 2 produce a rating distribution that is wider than 50%-chess. In particular, poker clearly fails to pass the 50% benchmark. Our calibration suggests that poker is roughly like 25%-chess.

Result 1 poses a puzzle. If poker is a game that depends predominantly on chance, then why are there poker professionals? It is undisputed that there are quite a number of professional poker players. Some are very well known from

TV shows and live events and may derive a substantial part of their income from advertisements. However, there are also numerous unknown professionals who make a living, in particular from online poker. These players continuously win more money than they lose, at least when their results are aggregated over longer time periods.[20] On first view, this might seem to be in conflict with our findings. However, there are two reasons why there is no contradiction. First, as we will show below, although the influence of skill in poker may be smaller than in other games, it is still significant. Online poker professionals often play many hours per day and several matches in parallel. Thus, by the sheer number of games, they can make a decent return despite being only marginally favored in each match. Second, game selection is an important factor in poker. This is a crucial difference about this between chess and poker. While in chess, one generally tries to find opponents of similar, or even slightly higher strength, in poker one tries the find an opponent who is as bad as possible (a "fish" in poker terminology). This becomes apparent when considering the mean rating differences of players who enter a match in Table 2. For all games except poker, the magnitude of this number is between one and two standard deviations of the rating distribution.[21] Poker, on the contrary, shows a value that corresponds to five times its standard deviation. Thus, it seems that professional online poker players can make a living by playing many games and by using additional information to identify weak players.[22]

In order to demonstrate that skill is important in the games we consider, we present the results of regressions which were inspired by the approach taken by

[20] One of the authors made this experience himself when he played poker to finance his studies.

[21] The players of the online browser games might also have an incentive to try to pick weak opponents, but a match-making algorithm does not allow them to choose freely. In addition, signals of player strengths are common knowledge, which means that potentially weaker players would be aware of entering an unfavourable competition (and can therefore avoid it).

[22] This information is not automatically available to every player. Statistics can be acquired through tracking software while playing, or a priori be purchased from special vendors. Generally, stronger players use these more often, leading to asymmetric information among players.

Regressions all players				Regressions regulars			
	#Obs	β_1	R^2**-value**		**#Obs**	β_1	R^2**-value**
Chess	8,274,204	0.373***	0.020	*Chess*	4,581,177	0.527***	0.016
Tetris	84,564	0.322***	0.024	*Tetris*	22,297	0.415***	0.012
50% Chess	8,274,204	0.252***	0.007	*Jewels*	425,091	0.402***	0.008
Jewels	844,934	0.267***	0.009	*50% Chess*	4,581,177	0.438***	0.007
Rummy	71,045	0.156***	0.005	*Rummy*	23,132	0.409***	0.008
Solitaire	1,248,696	0.158***	0.002	*Backgammon*	36,764	0.373***	0.007
Backgammon	80,031	0.124***	0.002	*Poker*	171,089	0.284***	0.002
Yahtzee	203,521	0.105***	0.001	*Solitaire*	826,629	0.251***	0.002
Crazy 8s	192,273	0.050***	0.000	*Yahtzee*	112,256	0.285***	0.003
Poker	329,258	0.087***	0.001	*Crazy 8s*	61,701	0.277***	0.003

*** $p < 0.001$

Table 4: Coefficients and R^2-values for Croson et al. regressions for all players and regulars

Croson, Fishman, and Pope (2008). Whenever a player competes in a match and has a history of matches played beforehand, we use his previous results to calculate his average performance in the past. Let $\bar{S}_i^{t-1}$ denote the average of all past scores of player i up to match $t-1$. Then, we estimate the effect of this previous average performance on the outcome of the current match.[23]

$$S_{ij}^t = \beta_0 + \beta_1 \cdot \bar{S}_i^{t-1} + \varepsilon_i^t$$

Whenever β_1 is significant and positive, we can conclude that skill plays a significant role. Furthermore, comparing across games, we can interpret a larger coefficient as a sign of more skill in a game.

We run the regressions with clustered standard errors on the player level. Table 4 shows the results for all players (left panel) as well as for regulars (right panel). The first thing to note is that the coefficient for past average rank is highly significant ($p < 0.001$) for all games we consider. As the past performance should

[23]The 50%-chess dataset uses a modified independent variable. The average performance in the past is based on half real, half random performances.

have no predictive power for future performance if the game in question is a game of pure chance, this suggests that for all of the games considered in Table 4 skill plays a significant role. We can thus confirm the results of earlier studies for poker, in particular, Croson et al. (2008). Remarkably, the coefficients in Table 4 have a very similar order as the one we obtained for our standard deviations measured using the best-fit Elo rating. To facilitate comparison, the games in Table 4 are presented in the same order as in Table 2.

Result 2: All games we consider (including poker) show a significant influence of skill.

5 Conclusion

The contribution of this paper is twofold. On the theoretical side we suggest a new way of classifying games as games of skill versus games of chance. Our preferred measure is the standard deviation of ratings after we rated all players according to a "best-fit" Elo rating. Most importantly, we provide a 50% benchmark that allows us to determine whether a game depends "predominantly" on chance. This benchmark is created by randomly replacing 50% of outcomes in our chess data set with coin flips. On the empirical side we employ large data sets from chess, poker, and online browser games to give our method a first practical test.

Our results clearly show that most popular two-player games in our data predominantly depend on chance in the sense that they did not pass the 50%-chess threshold. This holds in particular for poker, which we can classify as roughly "25%-chess". This does by no means imply that there is no skill in poker. However, if one adopts our view that "predominantly" is supposed to mean "by more than 50%", and if one accepts our way of inducing a 50%-benchmark, then, as a conclusion, poker is a game of chance.

Several points in our approach may be criticized. For example, one may argue that chess is not a game that consists to 100% of skill. However, as far as we

know, chess is universally accepted by courts as a game of skill. Furthermore, if we adopted a benchmark even stronger than chess, our 50%-benchmark would also move upwards, making our results conservative with respect to determining games of chance. Games that we classify as games that depend predominantly on chance when compared to 50%-chess would a fortiori be classified as such under a stricter benchmark.

One inevitable feature of any empirical approach is that our results depend on the population we observe. Suppose that we would observe chess matches in a completely homogenous population, i.e. when every player has exactly the same skill. If we applied our method to this sample, we would conclude that chess is a game of pure chance as the distribution of ratings would be very much concentrated at zero. Or consider a population in which players are completely "stratified", i.e. good players play only against good players and bad players only against bad players. This could happen because players are matched by the platform into very homogeneous groups (or because players choose similar opponents voluntarily). If the good players never play against the bad players, the best of the bad players will have a ranking comparable to the best of the best players (because they both win most of their games). As a result, the overall ranking distribution would be compressed. The Elo rating is capable of handling this issue if at least sometimes some of the good players are matched against some of the bad players. Transitivity of the Elo ranking will then detect the heterogeneity in skills, which allows it to rank the players accurately. For this reason, any ranking method that does not control for the strength of the opponents would underestimate the skill distribution.

The purpose of this paper is not to discuss the reasonableness of the current regulation of gaming. We leave open whether games that "predominantly depend on chance" *should* be treated differently from skill games. The legal status of gaming simply serves as a starting point for our analysis. However, we conjecture that games with a higher degree of chance elements might be more subject to problem gambling. It seems fair to assume that few people become addicted to playing chess for money, since repeated, predictable, losses against better players would

be hinder the addition. In poker, on the other hand, even a fairly inexperienced player may win a few hands or even a tournament and very good players may lose early. Given evidence on overconfidence (see e.g. Park and Santos-Pinto (2010)), this may lead to problem gambling.

Finally, we should point out that this paper is only a first step towards a broader research agenda. One limitation is that it applies only to two-player games. In future research we want to generalize the Elo rating to n-player games and conduct an analysis similar to the current one. The current rating can also be applied to other games or sports, if sufficient data is available.

6 Appendix

6.1 Description of browser games

We selected browser games for our analysis that do not differ significantly from popular versions of those games. Nevertheless, some adjustments were made by the providing website. On the one hand, they facilitate competitive matches in games that are originally single person games, on the other hand they balance the influence of random devices in order to allow for strategic gaming and "fairer" comparison of competitors.

The implementations of crazy eights as well as rummy do not differ much from the popular variants. Crazy eights (also known as "Mau-Mau") is a shedding-type card game with the objective to get rid of all cards. Rummy is a matching card game. Its' objective is to build melds and to get rid of all cards by doing so.

The two-player board game backgammon provided by the website is nearly identical to the popular version of the game. The goal for each player is to remove all of his playing pieces from the board.

The single player games solitaire (also known as "patience"), jewels and tetris each are complemented with a scoring scheme in order to establish a winner. In solitaire the players aim to sort a layout of cards. The initial setup of cards is

identical for both players in the online variant. Jewels and tetris are tile-matching puzzle games. While in jewels both players have to play the same patterns of gems, in tetris the order of tetrominos is predetermined and equal for the competitors. In all of these three games, identical strategies will lead to the exact same outcome.[24]

The latter also holds for the provided version of yahtzee (also known as "Kniffel"). It is a dice game with the objective to score by making certain combinations. All rolls are predetermined and identical for the players.

6.2 Minimization of loss function

Here we describe the numerical procedure used to minimize the quadratic loss function given in (1). Let

$$\mathcal{L}(k) := \frac{1}{T} \sum_{\substack{t \in T \\ i,j \in \rho(t)}} \left(S_{ij}^t - E_{ij}^t(k) \right)^2$$

be the value of the loss function for a given k-factor. The absolute loss by itself is not meaningful as it depends on the number of matches. Thus we will normalize the loss by considering the improvement relative to $\mathcal{L}(0)$, which is the loss when all ratings are set to the initial value of zero. For all games we considered, the loss value is roughly U-shaped, starting high at $\mathcal{L}(0)$ but increasing again after k^*. As an example see Figure 2, which shows the loss for the game of backgammon.

To find the minimum we conduct a grid search moving to a finer and finer grid in each iteration. We start by considering five equidistant k-values of 0, 20, 40, 60, and 80.[25] Suppose 20 produces the lowest loss among those five, then we continue by halving the grid size taking 20 as center point, i.e., the new grid will consist of 0, 10, 20, 30, and 40.

[24]Nevertheless, draws are very unlikely to occur, as time also counts towards the score and therefore an identical strategy would have to be identical in timing as well.

[25]We chose these initial values conservatively to guarantee that the solution to our minimization problem is in the interior of this interval.

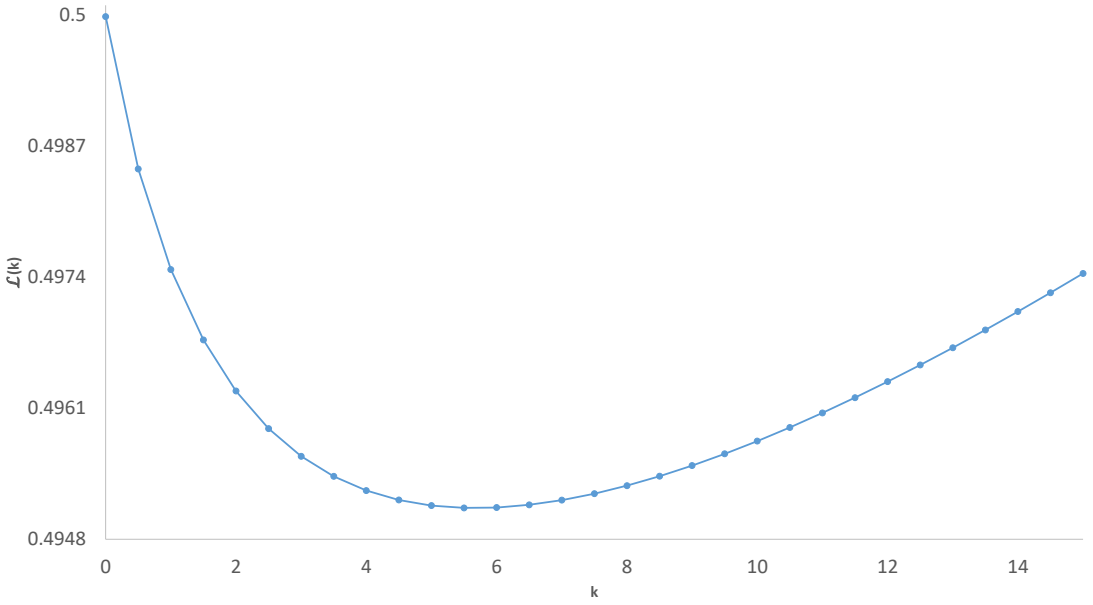

Figure 2: Loss as function of k-factor for the game "Backgammon"

We stop this procedure at k^* once we have achieved a desired degree of precision, which we define as

$$\frac{[\mathcal{L}(k_+) - \mathcal{L}(k^*)] + [\mathcal{L}(k_-) - \mathcal{L}(k^*)]}{\mathcal{L}(0) - \mathcal{L}(k^*)} < 10^{-6},$$

where k_+ denotes the grid point above k^* and k_- the grid point below k^* (see Figure 3).

Table 5 shows the results of the procedure for each dataset. It includes the optimal k-factor derived through the numerical algorithm, as well as the resulting value of the loss function when using this k-factor.[26] The value of the loss function can be interpreted similar to the Brier score (Brier, 1950). The lower this value, the more accurate are the predictions of outcomes. The value of 0.5 can be taken as benchmark, as this loss would result when predicting both players to be equally likely to win in each of the matches. In general, the accuracy of predictions seem to be correlated to the heterogeneity of skill within a game. On the other hand,

[26]For the datasets of chess and 50%-chess, we additionally list the value of the loss function when excluding draws. These values are more adequate to compare to those of the other games.

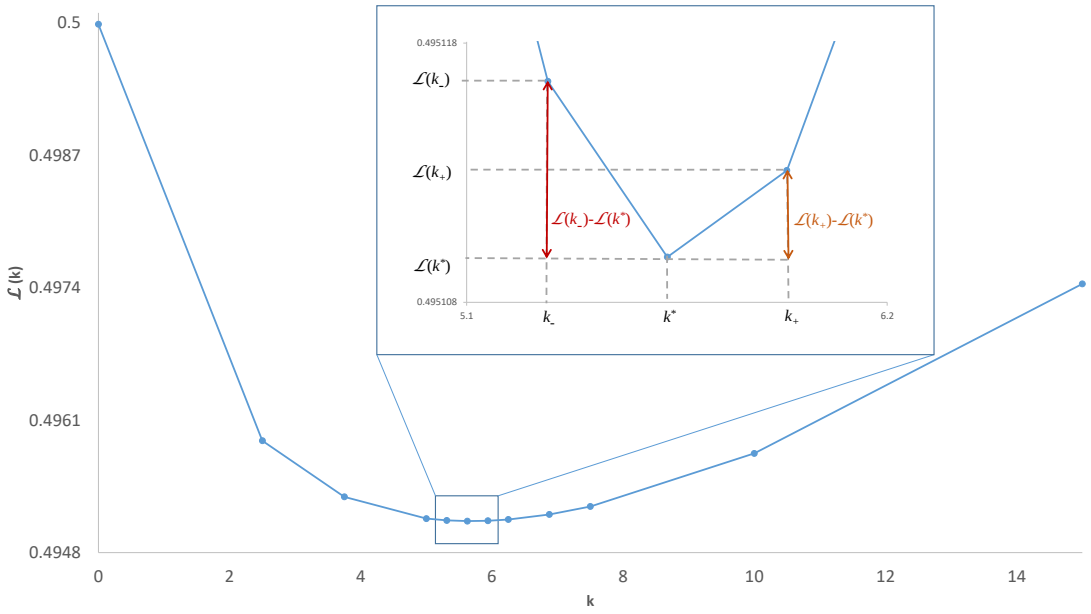

Figure 3: Numerical procedure for the game "Backgammon"

the value also depends on the competitions that one observes. Even with perfect approximation of playing strengths, it is necessary that players have a different level of skill in order to generate a lower loss contribution. From Table 2, we already know that the mean difference in ratings is considerably larger in the observed poker matches. Consequently, the calibrated poker ratings result in a lower loss value than some of the online games that show a larger heterogeneity of skill.

	Chess	*Tetris*	*50%-chess*	*Jewels*	*Rummy*	*Solitaire*	*Backgammon*	*Yahtzee*	*Crazy eights*	*Poker*
k^*	57.0	39.3	11.7	12.4	9.8	4.9	5.7	4.4	3.7	4.9
$\mathcal{L}(k^*)$	0.298	0.475	0.361	0.491	0.491	0.497	0.495	0.496	0.498	0.494
$\mathcal{L}(k^*)$ (no draws)	0.395		0.489							

Table 5: Derived k-factors and corresponding loss-function values

References

Borm, P., and B. van der Genugten (2001): "On a relative measure of skill for games with chance elements," *Top*, 9(1), 91–114.

Brier, G. W. (1950): "Verification of forecasts expressed in terms of probability," *Monthey Weather Review*, 78(1), 1–3.

Cabot, A., and R. Hannum (2005): "Poker: Public policy, law, mathematics, and the future of an American tradition," *TM Cooley L. Rev.*, 22, 443.

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(4), 25–28.

DeDonno, M. A., and D. K. Detterman (2008): "Poker is a skill," *Gaming Law Review*, 12(1), 31–36.

Dreef, M., P. Borm, and B. van der Genugten (2003): "On strategy and relative skill in poker," *International Game Theory Review*, 5(02), 83–103.

——— (2004a): "Measuring skill in games: Several approaches discussed," *Mathematical Methods of Operations Research*, 59.3, 375–391.

——— (2004b): "A new relative skill measure for games with chance elements," *Managerial and Decision Economics*, 25(5), 255–264.

Economist (2010): "You bet," July 8, 2010, 14–15.

Elo, A. E. (1978): *The rating of chessplayers, past and present*. Arco Pub., New York.

Fiedler, I. C., and J.-P. Rock (2009): "Quantifying skill in games - theory and empirical evidence for poker," *Gaming Law Review and Economics*, 13(1), 50–57.

Glickman, M. E. (1995): "A comprehensive guide to chess ratings," *American Chess Journal*, 3, 59–102.

Glickman, M. E., and T. Doan (2017): "The US chess rating system," *US Chess Federation*.

Kelly, J. M., Z. Dhar, and T. Verbiest (2007): "Poker and the law: is it a game of skill or chance and legally does it matter?," *Gaming Law Review*, 11.3, 190–202.

Larkey, P., J. B. Kadane, R. Austin, and S. Zamir (1997): "Skill in games," *Management Science*, 43(5), 596–609.

Levitt, S. D., and T. J. Miles (2014): "The role of skill versus luck in poker: Evidence from the world series of poker," *Journal of Sports Economics*, 15(1), 31–44.

Park, Y. J., and L. Santos-Pinto (2010): "Overconfidence in tournaments: Evidence from the field," *Theory and Decision*, 69(1), 143–166.

Siler, K. (2010): "Social and psychological challenges of poker," *Journal of Gambling Studies*, 26(3), 401–420.

van der Genugten, B., and P. Borm (2016): "Texas Hold'em: A game of skill," *International Game Theory Review*, 18(03), 1650005.

van Loon, R. J. P., M. J. van den Assem, and D. van Dolder (2015): "Beyond chance? The persistence of performance in online poker," *PLoS one*, 10(3), e0115479.