European Economic Review 127 (2020) 103472

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/euroecorev

Measuring skill and chance in games

Peter Duersch[1], Marco Lambrecht[1], Joerg Oechssler[1,*]

University of Heidelberg, Department of Economics, Bergheimer Str. 58, Heidelberg D-69115, Germany

ARTICLE INFO

Article history:
Received 4 December 2019
Revised 23 April 2020
Accepted 3 May 2020
Available online 19 May 2020

JEL classification:
K23
L83
C72

Keywords:
Elo
Ranking
Games of skill
Games of chance
Chess
Poker

ABSTRACT

Online and offline gaming has become a multi-billion dollar industry, yet, games of chance (in contrast to games of skill) 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 depend on chance, as this criterion is often used by courts. We apply the method to large datasets of various games (e.g. chess, poker, backgammon). 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 75% of the chess results are replaced by a coin flip.

1. Introduction

Online and offline gaming has become a multi-billion dollar industry. According to the Economist, the legal gambling market amounted to 335 billion US dollars 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 (Bewersdorff, 2004). This question has both legal and regulatory implications: 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 many countries winnings from games are treated differently for tax purposes if they are generated in games of skill rather than in games of chance.

So far, no universally accepted 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 clear guidance from the theoretical literature,

* Corresponding author.
E-mail addresses: duersch@xeeron.de (P. Duersch), marco.lambrecht@awi.uni-heidelberg.de (M. Lambrecht), oechssler@uni-hd.de (J. Oechssler).

[1] We like to thank the associate editor and two anonymous referees as well as the audiences at Heidelberg, Paris-Dauphine, ESA San Diego, GTM St Petersburg, HeiKaMaX, FU Berlin, RTG Mannheim/Heidelberg, Virginia Tech, City University of London, IMEBESS Florence, PCBS Prague, SING Bayreuth, ESA Berlin, Paderborn, Stony Brook, MIND games Pisa, ESEM Cologne, EALE Milan, LEG Tel Aviv, as well as Aleksandr Alekseev, Jörg Bewersdorff, Adam Dominiak, Mark Glickman, Fabian Krueger, Gerelt Tserenjigmid, Peter Winkler, and Rob van Zeijst for their helpful comments. Financial support by DFG grant OE 198/5-1 is gratefully acknowledged.

https://doi.org/10.1016/j.euroecorev.2020.103472
0014-2921/

courts and lawmakers 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 one specifies what predominantly means, for instance "more than 50 percent", the question remains, "50 percent of what?"

We propose a new method for measuring the skill and chance components of games and apply it to poker, chess, tennis, backgammon, and several other popular games. The main objective is to provide a method that allows to define a clear 50%-benchmark for the predominance of chance versus skill. Furthermore, the measure should be easily applicable to a variety of games and not specific to, for example, one particular type of poker. Our approach is empirical and benefits from the availability of very large datasets. 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 databases. We also have access to millions of observations from one of Europe's largest online gaming websites, which offers a variety of different games.

Our method can be described in two steps. First, we propose a measure for skill and chance in games. Then, we use this measure to define a 50%-benchmark. For the first step, rather than using performance measures like prize money won or finishing in the top x percentile in a tournament, we apply a complete rating system for all players in our datasets. In particular, we build on 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 is applicable to two player games immediately, and can be generalized to handle multiplayer competitions. We calibrate the Elo rating system to obtain a best fit for each game and type of competition individually.

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 we can predict the winner of a match. 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 our data, the rating distributions of all games are unimodal, which makes it possible to interpret the standard deviation of ratings as a measure of skill. Accordingly, the standard deviation is high in games of pure skill and with 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 performance. In this case, the standard deviation of ratings tends to zero.

In the second step, we propose an explicit 50%-benchmark for skill versus luck. We do this by constructing a hybrid game that is arguably exactly half pure chance and half pure skill. For the pure skill part we use chess as a widely accepted game of skill with the added benefit that there is an abundance of chess data. We construct our hybrid game by randomly replacing 50% of matches in the chess dataset 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 then compare the standard deviations of ratings for all of our games to 50%-chess as a benchmark.

One may argue that even chess contains an element of chance. For comparison, we also provide a more extreme benchmark. This "50%-deterministic" game consists of matches where the better player wins 50% of the matches right away, while the other 50% are decided by chance.

Applying our method to the data, we obtain a distribution of ratings for each game. As expected, chess and Go, as well as a traditional sport like tennis, have high standard deviations. Poker, on the other hand, has one of the narrowest distributions of all games. When we compare the games to our benchmarks, we find that poker, backgammon, and other popular online games are below the threshold of 50%-chess (and therefore also below the higher threshold of 50%-deterministic) 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 75% of the chess results are replaced by a coin flip. Furthermore, the amount of skill we find in poker is comparable to that of a deterministic game when 85% of the results are replaced by chance.

There are a number of earlier approaches in the literature that are mostly concerned with poker.[3] While most conclude that skill plays a statistically significant role in poker (a result we do not dispute at all), they do not provide a benchmark for classification. One interesting approach is to compare poker to sports or financial markets. Croson et al., 2008 compare data from poker to data from golf and find that past performances have about the same predictive power in both games. However, when we compare poker to tennis, we find large differences. 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).

[2] For example, 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 as one whose "outcome depends largely or wholly on chance" (translated by the authors, §3 Abs. 1 GlueStV).

[3] There has been an extensive debate in courtrooms as to whether poker is a game of chance or rather a game of skill. Different 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 (UIGEA), see Rose (2011). On the other hand, in 2012 a federal judge in New York ruled that poker is rather a game of skill, see *USA vs Lawrence Dicristina, US District Court Eastern District of New York, 11-CR-414*. Similarly, in other jurisdictions like e.g. Austria, Israel, and Russia, poker is categorized as a game of skill (Kelly et al., 2007). In Germany, courts still refer to a decision by the Reichsgericht from 1906 that considered poker as a game of chance (Holznagel, 2008).

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 et al. (2003, 2004a, 2004b), 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 approach is, however, limited to simplified versions of poker. Nevertheless, even for simple poker variants, the different studies report a substantial degree of skill. Van Essen and Wooders (2015) compare the behavior of online poker players to the behavior of novices for a stylized version of the game and find significant differences.

Larkey et al. (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, Potter van Loon et al., 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 be expressed in terms of a 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

Our empirical approach to measuring skill and chance is based on 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 players were 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 immediately applicable to two-player games. Furthermore, we introduce a generalization of the Elo rating for multiplayer competitions, which is based on the rank-ordered logit model (see e.g. Beggs et al., 1981).

Thus, the first step in our procedure is to rank all players in all games according to the Elo rating formula. This formula has one parameter that needs to be calibrated for each game. In Subsection 2.3 we explain in detail how this is done. Once all players are ranked, 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. Differences in Elo ratings of players correspond to their predicted winning probabilities via a logistic function.[4] Therefore, the heterogeneity of ratings is correlated to the predictability of outcomes and is a proxy for the amount of skill involved. The standard deviation of a well-calibrated rating should approach zero in a game of pure chance.[5] In a game of pure skill like chess, the standard deviation is very high.[6]

The standard deviations of ratings give us an ordinal measure as they allow us to make statements such as "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 allows to specify whether a game is "predominantly" a game of skill or chance, respectively. For this purpose our innovation is to construct a hybrid game that is a convex mixture of a game of pure skill and a coin flip. Chess is commonly regarded as an archetypical game of skill. It is also widely known and very large datasets are available, making it a good benchmark. Additionaly, we consider a more extreme theoretic benchmark for pure skill: a simulated game where all players can be ordered according to their playing strengths, and whenever two players face in a competition, the better one will always win. We call it "deterministic game". A coin flip, on the other hand, is an archetypical game of chance. We construct our hybrid games "x%-chess" and "x%-deterministic" by replacing randomly $(100-x)$% of matches in our data by a coin flip. In fact, for "x%-chess", 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 dataset, by a "win" with probability $\frac{1}{2}(1-\gamma)$ and a "loss" with probability $\frac{1}{2}(1-\gamma)$. For "x%-deterministic", we simulate a dataset of players with distinct strengths.[7] For every match, there is a chance of $(100-x)$% that the better player wins, while there is an x% chance that a coin flip will decide the winner.[8]

[4] While Elo's original proposal (Elo, 1978) was based on a normal distribution, a logistic one is used today by some chess federations.

[5] We confirm this by replacing all outcomes by random results in three datasets. The standard deviations take values of between 0.3 for the smallest to 0.003 for the largest dataset. For details, see Section A.5.

[6] In fact, purely deterministic outcomes would correspond to an infinite rating difference between any two matched players.

[7] In fact, we simulated datasets with different parameters (number of players in the dataset, average number of matches per player) and found that our analysis is robust to these changes. We provide the do-files of our simulations with the supplementary material of this paper.

[8] In this case, the coin does not allow a draw.

We will use "50%-chess" and "50%-deterministic" as our benchmarks since this seems to be the most plausible 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 and 50%-deterministic, we will say that the game is predominantly skill. If it is below, it is categorized as a game of predominantly chance. Yet, our method is quite flexible as it can be used to calculate any arbitrary version of x%-chess or x%-deterministic as a potential benchmark. Similarly, one could replace chess with any other pure skill game, e.g. Go.

2.1. Starting from two player 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 $i, j \in I$.[10] 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 $\rho(t)$.[11]

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 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

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

Expected scores range from zero (sure loss) to one (sure win). The parameter 400 in the logit function is a normalization used by chess federations which we retain for familiarity.[12] Given this parameter, a rating difference of 100 translates into an expected score of 0.64.

We normalize the initial rating of each player to $R_i^0 = 0$.[13] The Elo ratings of the players who were involved in match t are updated as follows,

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

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

2.2. Generalizing Elo ratings for multiplayer competitions

The concept of the Elo rating can be generalized to deal with multiplayer competitions by defining a proper framework. We adopt the perspective used by choice theorists, who extend models for pairwise comparison to n alternatives by applying a rank ordered logit model. Most of our notation of the two player case can be adopted, but some adjustments have to be made. We formalize this as follows.

Again, we have a finite set of players I to be ranked, a finite number of matches T, a finite series of outcomes from each match $t \in \{1, ..., T\}$, and the set of players involved in match $t \in T$ denoted by $\rho(t)$, where $\rho(t) \subset I$. Let n^t denote the number of players participating in t and let $S_i^t \geq 0$ be the outcome (payoff) in match t for player $i \in \rho(t)$.

After each match, players in $\rho(t)$ are ranked according to their performance. There are $n^t!$ different rankings. Let Q^t be the set of possible rankings in match t and $q^t \in Q^t$ a ranking. Then, q_k^t denotes the player ranked at position k (by ranking q^t) and $q^t(i)$ is the rank order of player $i \in \rho(t)$ under ranking q^t.

Each match t is characterized by a prize money structure that assigns a prize π_k^t to each position k in the ranking. We denote the probability of player i being ranked k-th in match t by $P^t(q^t(i) = k)$. Thus, the expected payoff of player i in match t equals

$$E_i^t := \sum_{k=1}^{n^t} \pi_k^t \cdot P^t(q^t(i) = k).$$

The rating R_i^t of player i after match t is an empirical measure of player i's playing strength. We assume a standard rank-ordered logit model (see e.g. Beggs et al., 1981). Thus the probability of a particular ranking q^t, where ranking position k ranges from 1 (first) to n^t (last) is

$$P(q^t) := \prod_{l=1}^{n^t - 1} \frac{e^{R_l^t}}{\sum_{j=l}^{n^t} e^{R_j^t}}.$$

[9] cf. footnote [1].

[10] Matches are ordered chronologically by start time.

[11] This, for now, is always a pair of players.

[12] Due to our calibration method, the use of this parameter is without loss of generality (Appendix A.3 shows this formally).

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

We can use this to calculate the probability $P^t(q^t(i)=k)$ of player i ending up at a given position k. It is the sum of probabilities of rankings q^t in which the k-th ranked player is player i.

$$P^t\left(q^t(i)=k\right)=\sum_{q^t\in\left\{q^t\in Q^t\,\middle|\,q^t_k=i\right\}}P\left(q^t\right)$$

Let $\pi^t_{\max}$ denote the maximum possible payoff of match t. We normalize the outcome as well as its expected value to represent shares of this payoff,

$$\hat{S}^t_i=\frac{1}{\pi^t_{\max}}\cdot S^t_i,\qquad \hat{E}^t_i=\frac{1}{\pi^t_{\max}}\cdot E^t_i.$$

Due to the normalization, these shares range from zero to one.[14]

Once more, the initial rating of each player is set to $R^0_i=0$. Subsequently, the Elo ratings of the players who were involved in match t are updated,

$$R^{t+1}_i=R^t_i+k\cdot(\hat{S}^t_i-\hat{E}^t_i),\ \forall\, i\in\rho(t),$$

and the ratings of players who are not involved in match t do not change, i.e. $\forall i\notin\rho(t): R^{t+1}_i=R^t_i$.

Before we measure the standard deviation, we multiply all ratings by 400 to retain comparability with our two player ratings and established chess ratings.

2.3. Calibrating the Elo ratings

While the actual scores S^t_i are observed in our data, the expected scores are determined recursively and depend on k. To indicate this we write $E^t_i(k)$. 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 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.[15]

Our approach is to calibrate a 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:

$$k^*:=\arg\min_k\frac{1}{T}\sum_{\substack{t\in T\\ i\in\rho(t)}}\left(S^t_i-E^t_i(k)\right)^2. \tag{2}$$

The graph of this loss function is roughly U-shaped for all of our datasets, and we derive the solution to the minimization problem numerically.[16] Note that for a game of pure chance, k^* takes a value very close to zero, leading to nearly identical ratings for every player (independent of the number of observations).

It may be tempting to interpret a high k^*-factor as a sign of a game of skill. 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 all 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 games. These include matches of chess, poker, and online browser games. We remove matches between isolated players (i.e., players who are not connected to the main dataset via playing).[17] The remaining datasets are still quite large and are summarized by the statistics in Table 1. For each game, Table 1 lists the total number of matches, as well as the number of players, and the number of "regulars". The latter are those players who play at least 25 matches within our data. Furthermore, we report the maximum number of matches played by a single player.

[14] Note that the sum of these scores equals the sum of all payoffs divided by the maximum payoff. This sum amounts to 1 if and only if the maximum payoff is the only payoff, i.e. a "winner-takes-all " competition.

[15] 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.

[16] See Appendix A.4 for an example graph as well as an exact description of the numerical procedure.

[17] This removes, for most games, less than 0.3% of the original datasets. For poker and Go however, about 1.3% of the data had to be removed.

Table 1
Statistics on the number of players, matches and regulars in the datasets. Number of players per match in parentheses for games with different formats.

	#Players	#Regulars	#Matches	Max. Matches
Chess	233,683	71,345	4,253,630	2,280
Poker (9p)	105,787	5,799	94,261	3,095
Poker (2p)	55,158	1,883	191,704	7,531
Jewels (2p)	38,878	7,770	441,905	1,649
Poker (6p)	38,277	966	26,975	1,100
Solitaire (2p)	33,762	9,374	641,220	3,277
Go	25,888	3,165	222,334	3,722
Tennis	21,034	6,502	614,714	1,243
Jewels (5p)	19,923	5,664	154,311	1,713
Solitaire (3p)	17,240	4,355	200,489	6,980
Solitaire (5p)	15,747	4,796	150,084	3,322
Yahtzee (4p)	12,760	3,535	134,455	3,649
Crazy 8s (2p)	12,392	2,136	102,187	2,945
Tetris (2p)	10,484	882	47,507	514
Yahtzee (2p)	9,969	1,678	106,722	1,378
Yahtzee (3p)	9,932	1,467	61,212	1,847
Skat	8,123	793	28,262	571
Rummy (2p)	7,719	672	39,349	3,026
Crazy 8s (3p)	6,872	190	11,062	656
Backgammon	4,229	780	42,126	1,301
Tetris (3p)	2,926	340	11,590	620

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.[18] The resulting subset consists of roughly 4.25 million matches from more than 230,000 players.

The poker data consist of *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. Each tournament is treated as one match.[19] We purchased the data from "HH Smithy", a commercial provider of poker hand histories. The data we use for this project include 55,158 players who participate in 191,704 tournaments for the two-player version (so-called "heads-up"). Furthermore, we analyse 26,975 tournaments of 6 player poker ("short-handed") including 38,277 players, as well as 94,261 tournaments of 9 player poker ("full-ring") including 105,787 players. All of these tournaments are "No Limit Texas Hold'em" matches, which is the most popular type of poker online. They took place between February 2015 and February 2017. The entry fee for each tournament was $3.50.

For Go data, we use a database dump from the popular webpage online-go.com published on Github.[20] We restrict attention to 19x19 Go games played without handicap, with Komi 6.5 and using Japanese rules. This corresponds to the default used on the website for creating a new game. In total, the data used consist of 222,334 matches played by 25,888 players.

In addition, we received 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 games that are (more or less) well-known, or comparable to well-known games, giving us more than 2 million matches. The number of different players for each game range from about 3,000 to 40,000. The games used are online versions of rummy, tetris, backgammon, skat, jewels, solitaire, yahtzee, and crazy eights.[21]

Most people would consider sports as games of skill. It may therefore be useful to compare our games also against a suitable two-person sport. Thus, we use a large database on men's tennis, which was collected by Jeff Sackmann.[22] The 614,714 matches we analyze were played at Grand Slam tournaments, ATP World Tour, ATP Challenger Tour, and ITF Future tournaments by 21,034 male tennis players between 1968 and 2017.[23]

[18] 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.

[19] Unlike cash-game poker, the participants of these heads-up tournaments cannot leave the match after every single hand (unless they choose to give up). They commit to this when starting the tournament. Therefore, we treat each heads-up tournament as one "match" in the sense of Section 2.

[20] https://github.com/gto76/online-go-games using commit 9fe78d5 from 6 Mar 2016.

[21] For a detailed description of these games, see Appendix A.2.

[22] https://github.com/JeffSackmann

[23] The top tennis players in the world compete in Grand Slam tournaments and the ATP World Tour, while the ATP Challenger Tour is considered to be the second highest level of competition. The players on this tour, among them talented young players, try to qualify for the ATP World tour. The ITF Future tournaments finally are the lowest tier of professional tennis and the starting point for nearly every professional player.

Table 2
Summary statistics of rating distributions - regulars only

	Std. Dev.	Min.	1%	99%	Max.	p^{sd}	p_1^{99}	Rep.
Go	278.9	-659.2	-440.1	859.3	1,432.2	83.3	99.9	1
Tennis	218.8	-389.6	-242.9	790.4	1,438.0	77.9	99.7	1
Chess	171.7	-684.6	-289.4	579.6	945.3	72.9	99.3	3
50% Determ.	122.8	-304.7	-263.0	244.1	295.7	67.0	94.9	5
Tetris (2p)	95.7	-372.0	-221.2	269.5	374.8	63.4	94.4	7
Tetris (3p)	61.6	-143.0	-116.8	204.3	256.7	58.8	86.4	15
Jewels (2p)	48.8	-411.1	-154.7	109.5	225.0	57.0	82.1	23
50% Chess	44.9	-201.7	-75.5	159.0	324.4	56.4	79.4	27
Rummy (2p)	35.9	-137.3	-64.5	103.0	121.7	55.1	72.4	43
Solitaire (5p)	32.7	-142.3	-75.0	99.9	218.6	54.7	73.2	51
Skat	29.4	-96.7	-58.4	89.1	132.5	54.2	70.0	65
Backgammon	24.8	-120.1	-62.9	73.2	130.1	53.6	68.6	89
Solitaire (2p)	24.5	-176.8	-60.8	63.9	122.5	53.5	67.2	93
Poker (6p)	23.2	-69.3	-45.6	71.2	86.5	53.3	66.2	103
Poker (2p)	22.9	-98.6	-40.9	81.1	123.1	53.3	66.9	105
Jewels (5p)	22.0	-225.5	-53.3	57.9	275.9	53.2	65.5	115
Yahtzee (2p)	20.8	-65.3	-46.1	65.3	86.3	53.0	65.5	127
Poker (9p)	18.3	-159.7	-37.5	58.1	93.7	52.6	63.4	165
Solitaire (3p)	18.2	-99.7	-46.1	54.3	108.9	52.6	64.1	165
Yahtzee (4p)	17.1	-86.3	-35.9	53.9	104.3	52.5	62.6	189
Yahtzee (3p)	16.8	-50.5	-34.5	57.9	93.1	52.4	63.0	195
Crazy 8s (2p)	15.2	-105.5	-32.0	39.9	184.8	52.2	60.2	239
Crazy 8s (3p)	2.3	-5.4	-4.6	9.6	9.9	50.3	52.0	12,637

4. Results

In this section, we present the results of our analysis, in particular, the standard deviations of the respective best-fit Elo rating distributions, which resulted from our rating procedure described in Section 2.3. Due to the fact that the Elo rating is based on an updating formula, the approximations of ratings become more and more meaningful the larger the number of matches played by a given player is. Thus, it seems prudent to require players to have played a certain minimum number of matches before including them into the rating distributions. On the other hand, when the required minimum number of matches is too high, we lose too many observations in some games. Since there is no obvious a priori cutoff value for the minimum number of matches, we calculate the rating distributions for all possible cutoffs from 1 to 100 games.

Figure 1 shows the standard deviations for all two player games depending on the cutoff number of matches. We split the set of games into two graphs with different scales on the y-axis. The top graph of Figure 1 shows the games with relatively high standard deviations, the bottom one those with relatively low standard deviations. Our 50%-chess benchmark is included in both graphs to facilitate interpretation. The benchmark 50%-deterministic is depicted in the top graph. Note that we derived a threshold value which is independent from the minimum number of matches per player and therefore present it as a straight line. The value is robust to variations of the parameters of our simulations.

Several facts are apparent from Figure 1. First, the standard deviations for all games show an upward trend when we increase the cutoff number of matches. Initially the increase is quite steep but flattens out quickly. More importantly, for cutoffs of 25 matches and higher there is hardly any change in the relative order of games, which is really the main focus of our analyis.[24] Thus, we pick a cutoff number of matches of 25 as our preferred version and call such players “regulars”. However, we also report tables with cutoffs of 1 and 100 in Appendix A.1.

In Table 2 we report summary statistics for the Elo rating distributions of regular players. 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 table according to this value. Via formula (1), we 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. We refer to this probability as p^{sd}. For comparison, we also provide the winning probablities when a 99% percentile player is matched against a 1% percentile player, which we call p_1^{99}. The winning probability p^{sd} can be used to calculate the number of matches necessary so 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%.[25] This number is reported in the repetitions column (abbreviated “Rep.”).

Table 2 confirms in more detail what was already apparent from Figure 1. The rating distributions of chess, Go, and tennis have very high standard deviations above 170. Our benchmark 50%-chess has a standard deviation of 44.9,[26] and almost all

[24] This also holds true for the respective graphs of the multiplayer games.
[25] This definition is used by e.g. Potter van Loon et al., 2015.
[26] Note that 50%-chess has a standard deviation that is substantially below 50% of the standard deviation of chess.

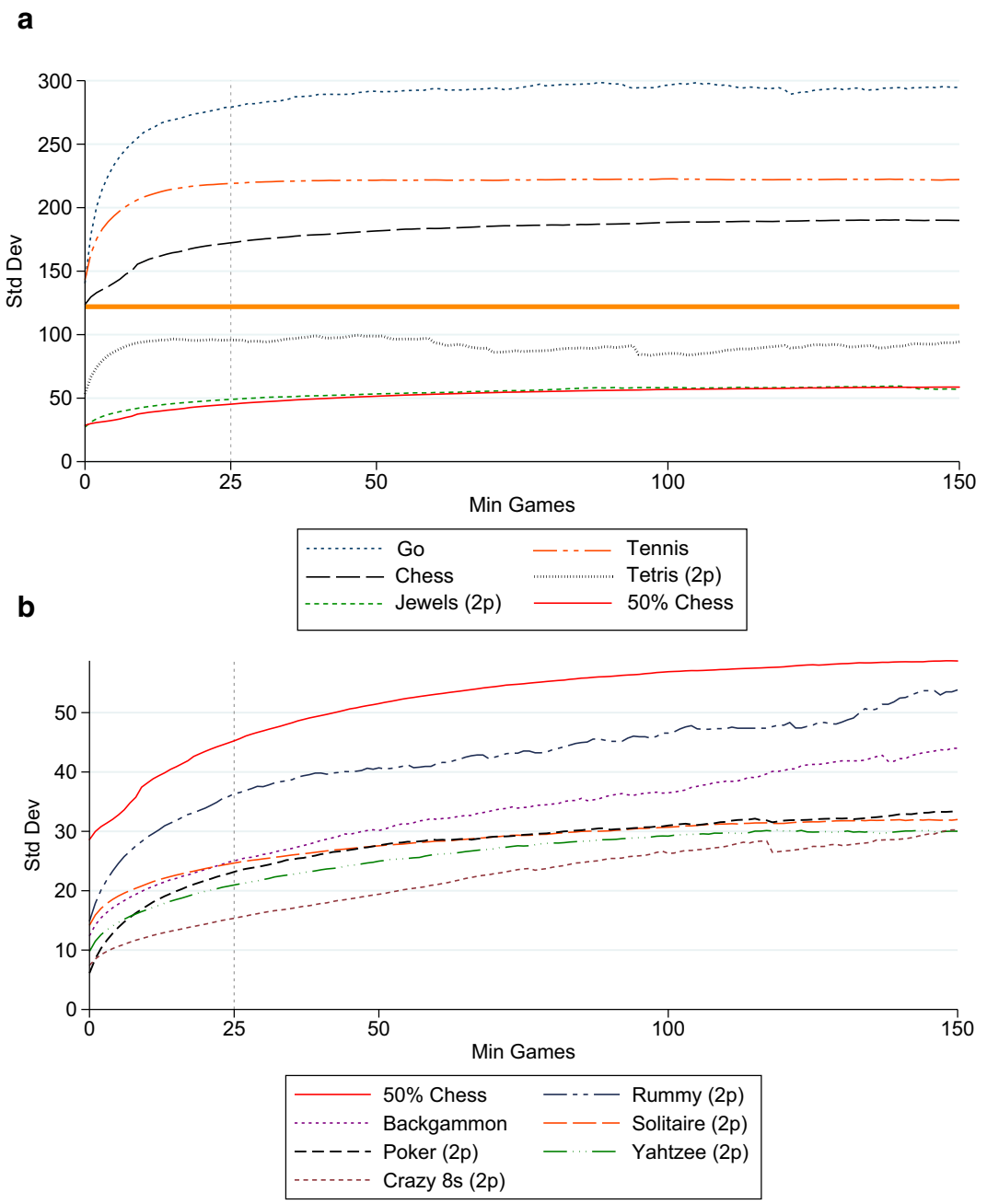

Fig. 1. Standard deviation of rating distributions for different cut-off values (min. number of matches per player). Note: The top graph depicts games with relatively high standard deviations, the bottom one those with relatively low standard deviations. The 50%-chess benchmark is included in both graphs to facilitate interpretation. The benchmark 50%-deterministic is depicted in the top graph as a thick solid line.

of the other games have a standard deviation substantially below this benchmark. Alternatively we could compare to the benchmark of 50%-deterministic, which would make our results even stronger in all cases. The online version of tetris is the single browser game that exhibits a larger heterogeneity of skill and positions itself clearly above the threshold of 50%-chess (but below 50%-deterministic) in all versions. Poker, on the other hand, is clearly below the threshold with a standard deviation of about 18-23 depending on the number of players, which ranks quite low in the list of all games we consider. Regarding winning probabilities, p^{sd}, a poker player who is one standard deviation better than his opponent seems to have not more than a 53.3% chance of winning the match. This translates into more than 100 repetitions that are needed for

Table 3
Average rating difference of pairs of matched players and standard deviations of ratings in two-player games

	Avg. rating diff in data	Std. Dev. of ratings	Ratio
Go	220.0	278.9	0.79
Tennis	155.1	218.8	0.71
Chess	112.8	171.7	0.66
Tetris (2p)	72.9	95.7	0.76
Jewels (2p)	39.3	48.8	0.80
50% Chess	36.3	44.9	0.81
Rummy (2p)	37.1	35.9	1.03
Backgammon	25.6	24.8	1.03
Solitaire (2p)	22.2	24.5	0.91
Poker (2p)	30.1	22.9	1.32
Yahtzee (2p)	22.4	20.8	1.08
Crazy 8s (2p)	12.7	15.2	0.84

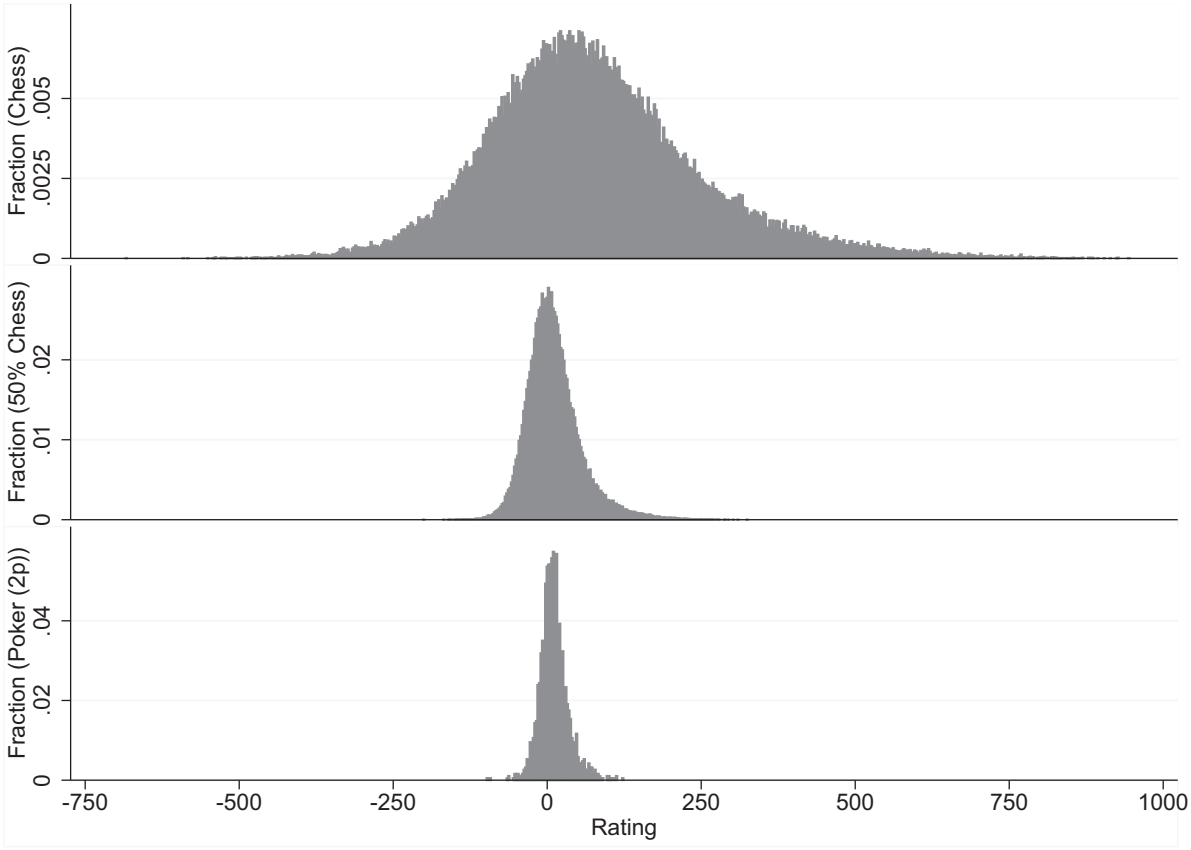

Fig. 2. Rating distributions for chess, 50%-chess, and poker (2p) - regulars

the better player to be ahead of his opponent with at least 75% probability. Similarly, the card game Skat fails to make the threshold of 50% skill.[27]

The histograms in Figure 2 provide the full distributions of Elo ratings of regulars for 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 with most of the ratings concentrated around 0.

We can now also reverse our procedure and ask: how much chance do we have to inject into chess and the deterministic game to obtain a distribution of player ratings similar to poker.[28] As a result, we find that we would have to replace roughly

[27] German courts refer to Skat as a game of skill, if it is played in tournaments and repeated for at least 36 times (see Bewersdorff (2004)). On the online platform, matches consisted of three to twelve repetitions.

[28] For details on this procedure, see Appendix A.6

Table 4
Coefficients, standard errors clustered on player level, t-values and R^2-values for regression specification (3)

	#Obs	β_1	Robust Std. Err.	t-value	R^2-value
Go	357,564	0.662***	0.008	86.81	0.058
Tennis	1,150,040	0.597***	0.006	97.55	0.026
Chess	7,082,266	0.436***	0.003	160.94	0.018
Tetris (2p)	56,204	0.331***	0.031	10.73	0.012
Jewels (2p)	706,626	0.306***	0.010	31.59	0.007
50%-Chess	7,082,266	0.331***	0.003	97.60	0.007
Rummy (2p)	48,207	0.270***	0.022	12.38	0.006
Backgammon	64,594	0.211***	0.023	9.07	0.003
Solitaire (2p)	1,133,111	0.200***	0.006	32.75	0.002
Poker (2p)	235,780	0.182***	0.015	12.26	0.001
Yahtzee (2p)	170,365	0.175***	0.014	12.57	0.002
Crazy 8s (2p)	145,472	0.103***	0.018	5.77	0.001

*** $p < 0.001$

3 out of 4 chess games by a coin flip in order to produce a rating distribution as the one in poker (i.e., the standard deviation of ratings is about the same in poker and 25%-chess). In our deterministic data, about 85% of the results have to be replaced by chance.

Result 1: Most of the games we consider produce rating distributions that are narrower than 50%-chess. In particular, poker clearly fails to pass the 50% benchmark. Our calibration suggests that poker is roughly like 25%-chess or 15%-deterministic.

Result 1 poses an empirical 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 of which are very well known from TV shows and live events. In addition, there are also numerous unknown professionals who seem to be able to 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.[29] 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 matches, 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 between chess and poker. In chess, one is mostly matched with opponents of similar strength.[30] On the other hand, in poker players try to find an opponent who is as bad as possible (a "fish" in poker terminology). This becomes apparent when considering the difference in playing strengths of the average pair of players entering a match in Table 3. For most two-player games, the magnitude of this difference is between 0.66 and 1 standard deviations of the rating distribution. Poker, on the contrary, shows a value that corresponds to 1.32 times its standard deviation. Thus, it seems that professional online poker players can make a living by playing many matches and by using additional information to identify weak players.[31]

In order to demonstrate that skill is important in the games we consider, we present the results of simple OLS regressions which are inspired by the approach taken by Croson et al., 2008. Whenever a player competes in a match, we use his previous results to calculate his average performance in the past.[32] 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.[33]

$$S_{ij}^t = \beta_0 + \beta_1 \cdot \bar{S}_i^{t-1} + \varepsilon_i^t \tag{3}$$

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

We run OLS regressions with standard errors clustered on the player level. Table 4 shows the results. The first thing to note is that the coefficients for past average rank are highly significant ($p < 0.001$) for all games we consider. As the past performance should 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 statistically 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

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

[30] Partially, this is due to players striving to find worthy opponents where no money is at stake, and partially due to tournament organizers enforcing it via the usage of swiss matching.

[31] 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.

[32] To be consistent with our previous approach, we analyze regulars, i.e. those players who made at least 25 games within the dataset.

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

similar order as the one we obtained for our standard deviations measured using the best-fit Elo rating, despite using a different methodology. 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 statistically 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 rank 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 dataset as well as 50% of outcomes in an artificial deterministic dataset with coin flips. On the empirical side we employ large datasets from chess, poker, Go, and online browser games to give our method a first practical test.

Our results clearly show that most popular games in our data predominantly depend on chance in the sense that they do not pass the 50% threshold. This holds in particular for poker, which we can classify as roughly "25%-chess" or "15%-deterministic". 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.

There are some caveats to mention. One may argue that "predominantly" might not translate to "more than 50%". Of course, politicians may decide to use a less strict benchmark instead, e.g. by changing the wording of the law to require less than predominance of skill. As long as a percentage-based interpretation is possible, our benchmark can easily be changed to accommodate "x%-chess" for any value of x.

It is worth noting that, when replacing chess or deterministic outcomes by a coin flip, the model and its transitivity are affected (e.g., for 50%-chess winning probabilities are bounded between 0.25 and 0.75, while extreme rating differences would imply winning probabilities close to 0 and 1). However, we do not want to rate any actual 50%-chess players but rather use this artificial game as a benchmark.[34]

One inevitable feature of any empirical approach is that our results depend on the population we observe.[35] Suppose that we observed chess matches in a completely homogeneous population, where every player had exactly the same skill. If we applied our method to this sample, we would conclude that chess is a game of chance as the distribution of ratings would be very much concentrated at zero. Or consider a population with separate pools of players such that players are completely "stratified", i.e. good players play only against good players and bad players only against bad players. This could happen when players are matched by a platform into extremely 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. Gambling regulation is a much-debated issue. While arguably pathological gambling (or problem gambling) imposes social costs on societies, both the identification of pathological gamblers as well as the estimation of the respective welfare losses are difficult. A recent study by Filippin et al. (2020) tries to identify pathological gamblers by items from the DSM-5 of the American Psychiatric Association. They find that gamblers showing severe scores of pathological gambling are more likely to play games that are pure chance games versus games that have some skill components. Similarly, Binde et al., 2017 find that the percentage of problem gamblers is higher among players who choose to play casino games than among players who prefer sports betting. In general, it seems fair to assume that few people become addicted to playing chess for money, since repeated, predictable, losses against better players would reduce the likelihood of addiction. 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. Park and Santos-Pinto (2010) find that overconfidence differs significantly for chess and poker players, which might contribute to games of skill being less problematic than those with a higher degree of randomness.[36]

In this study, we leave open whether games that "predominantly depend on chance" *should* be treated differently from skill games. We are neither challenging nor justifying the decision of legislators to have a binary classification. The legal status of gaming simply serves as a starting point for our analysis. However, we conjecture that games with a higher degree

[34] One could also adjust the method for 50%-chess and calculate expected scores according to $\hat{E}^t_{ij} = 0.5 \cdot 0.5 + 0.5 \cdot E^t_{ij}$, which is a combination of the expected outcome of a coinflip and the original Elo formula. Approximating the players' ratings with this adjusted formula leads (after calibration of the optimal k-factor) to $p^{sd} = 56.7$ and $p^{99}_1 = 72.4$, which would not change our main results.

[35] For our algorithm, it is sufficient to track the players and the winner of a competition. However, if you remove the tracking of players (i.e., "anonymous data"), the results of a game of pure skill (say, our "deterministic" game) would be indistinguishable from the results of a game of pure chance (i.e., coinflip results). For our research, we aimed to ensure that our datasets are as comprehensive and representative as possible. Unfortunately, no purely theoretical approach exists that is able to address games like poker in its full complexity.

[36] For more evidence on overconfidence, see also Camerer and Lovallo (1999), Biais et al. (2005), Malmendier and Tate, 2005, and Malmendier and Tate, 2008.

Table 5
Summary statistics of rating distributions - all players

	Std. Dev.	Obs.	Min.	1%	99%	Max.	p^{sd}	p_1^{99}	Rep.
Go	140.4	25,888	-659.2	-299.5	538.6	1,432.2	69.2	99.2	3
Tennis	143.0	21,034	-389.6	-212.3	597.3	1,438.0	69.5	99.1	3
Chess	123.8	233,683	-684.6	-247.8	440.6	945.3	67.1	98.1	5
50% Determ.	122.8	1,000	-304.7	-263.0	244.1	295.7	67.0	94.9	5
Tetris (2p)	53.4	10,484	-372.0	-123.4	184.2	374.8	57.6	85.5	19
Tetris (3p)	28.8	2,926	-143.0	-70.3	107.8	284.1	54.1	73.6	67
Jewels (2p)	27.2	38,878	-411.1	-79.1	80.1	225.0	53.9	71.4	75
50% Chess	28.6	233,683	-201.7	-59.0	110.9	324.4	54.1	72.7	67
Rummy (2p)	14.9	7,719	-137.3	-37.6	55.7	121.7	52.1	63.1	249
Solitaire (5p)	19.3	15,747	-142.3	-51.0	71.8	218.6	52.8	67.0	149
Skat	12.3	8,123	-96.7	-30.3	45.6	132.5	51.8	60.8	365
Backgammon	12.4	4,229	-120.1	-32.8	42.4	130.1	51.8	60.7	359
Solitaire (2p)	14.2	33,762	-176.8	-40.1	48.2	122.5	52.0	62.4	275
Poker (6p)	5.9	38,277	-69.3	-12.3	20.5	86.5	50.8	54.7	1,581
Poker (2p)	6.1	55,158	-98.6	-12.1	20.5	123.1	50.9	54.7	1,471
Jewels (5p)	12.6	19,923	-225.5	-38.5	40.4	275.9	51.8	61.2	345
Yahtzee (2p)	9.8	9,969	-65.3	-26.3	38.4	86.3	51.4	59.2	577
Poker (9p)	5.5	105,787	-159.7	-12.2	20.7	93.7	50.8	54.7	1,793
Solitaire (3p)	10.0	17,240	-99.7	-28.0	36.2	108.9	51.4	59.1	553
Yahtzee (4p)	9.6	12,760	-86.3	-24.1	38.2	104.3	51.4	58.9	601
Yahtzee (3p)	7.4	9,932	-50.5	-18.7	29.4	93.1	51.1	56.9	991
Crazy 8s (2p)	7.4	12,392	-105.5	-18.1	23.6	184.8	51.1	56.0	997
Crazy 8s (3p)	0.7	6,872	-5.4	-1.6	2.2	9.9	50.1	50.5	56,868

of chance elements might be more subject to problem gambling (and that games of chance potentially impose higher social costs than games of skill). In order to eventually analyze potential welfare losses from players choosing games of chance over games of skill, one first needs to define a validated measure of skill and chance. Thus, we see our study as an important building block for assessing the welfare cost of gambling in future studies.

Appendix A

A1. Results for all players and players with at least 100 matches in our dataset

As a robustness check, Table 5 shows results for the rating distributions when all players are included (i.e. when the cutoff for the minimum number of matches is set to 1). To facilitate comparison, the games are presented in the same order as in Table 2. It is noticeable that the standard deviations of all games are substantially lower than in Table 2. This is a consequence of the fact that our data include many players who only compete in a few matches and whose ratings therefore remain close to the initial value of 0. However, the overall ranking of games changes little. Go and poker are ranked slightly lower if all players are included.

Table 6 shows the results when the cutoff for the minimum number of matches is set to 100. When restricting the distributions to these players, the order of games is nearly the same as in Table 2. However, note that for some games the number of observations is very low. Poker is still clearly below the benchmark of 50%-chess and 50% deterministic.

Figures 3 and 4 show the histograms corresponding to Figure 2 when the cutoff for the minimum number of games is set to 1 or 100, respectively. Again, the qualitative results are independent of the cutoff chosen.

A2. Description of browser games

From the multitude of games that are offered on one of Europe's largest online gaming platform, we selected browser games that do not differ significantly from popular versions of those games. Nevertheless, some adjustments were made by the platform. On the one hand, games that are originally single person games are played as tournaments, on the other hand the platform tries to minimize the influence of random devices e.g. by giving competitors the same cards or dice rolls.

The implementations of skat, crazy eights and 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 with the objective to build melds and to get rid of all cards by doing so. Skat is a three-player card game that is specifically popular in Germany.

The two-player board game backgammon offered by the platform 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 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

Table 6
Summary statistics of rating distributions - players with 100 or more games only

	Std. Dev.	Obs.	Min.	1%	99%	Max.	p^{sd}	p_1^{99}	Rep.
Go	296.2	1,023	-659.2	-464.1	933.6	1,432.2	84.6	100.0	1
Tennis	222.6	3,419	-317.0	-156.9	877.1	1,438.0	78.3	99.7	1
Chess	188.3	18,963	-684.6	-200.9	703.4	945.3	74.7	99.5	3
50% Determ.	122.8	1,000	-304.7	-263.0	244.1	295.7	67.0	94.9	5
Tetris (2p)	84.7	139	-224.1	-156.1	260.3	308.2	62.0	91.7	9
Tetris (3p)	63.2	45	-121.2	-121.2	161.6	161.6	59.0	83.6	15
Jewels (2p)	58.2	1,899	-411.1	-205.5	128.8	225.0	58.3	87.3	17
50% Chess	56.8	18,963	-201.7	-73.5	205.6	324.4	58.1	83.3	17
Rummy (2p)	46.5	108	-137.3	-100.1	118.3	121.7	56.7	77.9	25
Solitaire (5p)	42.2	1,759	-142.3	-90.2	114.4	218.6	56.0	76.5	31
Skat	42.0	111	-96.7	-82.9	112.8	132.5	56.0	75.5	31
Backgammon	36.4	179	-120.1	-86.1	106.1	130.1	55.2	75.2	41
Solitaire (2p)	30.7	3,297	-176.8	-80.5	76.9	122.5	54.4	71.2	59
Poker (6p)	32.0	172	-69.3	-58.6	85.8	86.5	54.6	69.7	55
Poker (2p)	30.8	446	-98.6	-53.1	104.2	123.1	54.4	71.2	59
Jewels (5p)	30.6	1,951	-225.5	-67.5	76.1	275.9	54.4	69.6	59
Yahtzee (2p)	29.2	444	-65.3	-55.5	77.8	86.3	54.2	68.3	65
Poker (9p)	27.2	1,263	-159.7	-62.4	67.4	93.7	53.9	67.9	75
Solitaire (3p)	25.6	1,387	-99.7	-57.9	70.2	108.9	53.7	67.6	85
Yahtzee (4p)	23.7	1,322	-86.3	-47.2	69.4	104.3	53.4	66.2	99
Yahtzee (3p)	24.2	371	-50.5	-40.4	78.9	93.1	53.5	66.5	95
Crazy 8s (2p)	26.7	302	-105.5	-50.5	60.5	184.8	53.8	65.4	77
Crazy 8s (3p)	4.4	15	-5.4	-5.4	9.9	9.9	50.6	52.2	2,825

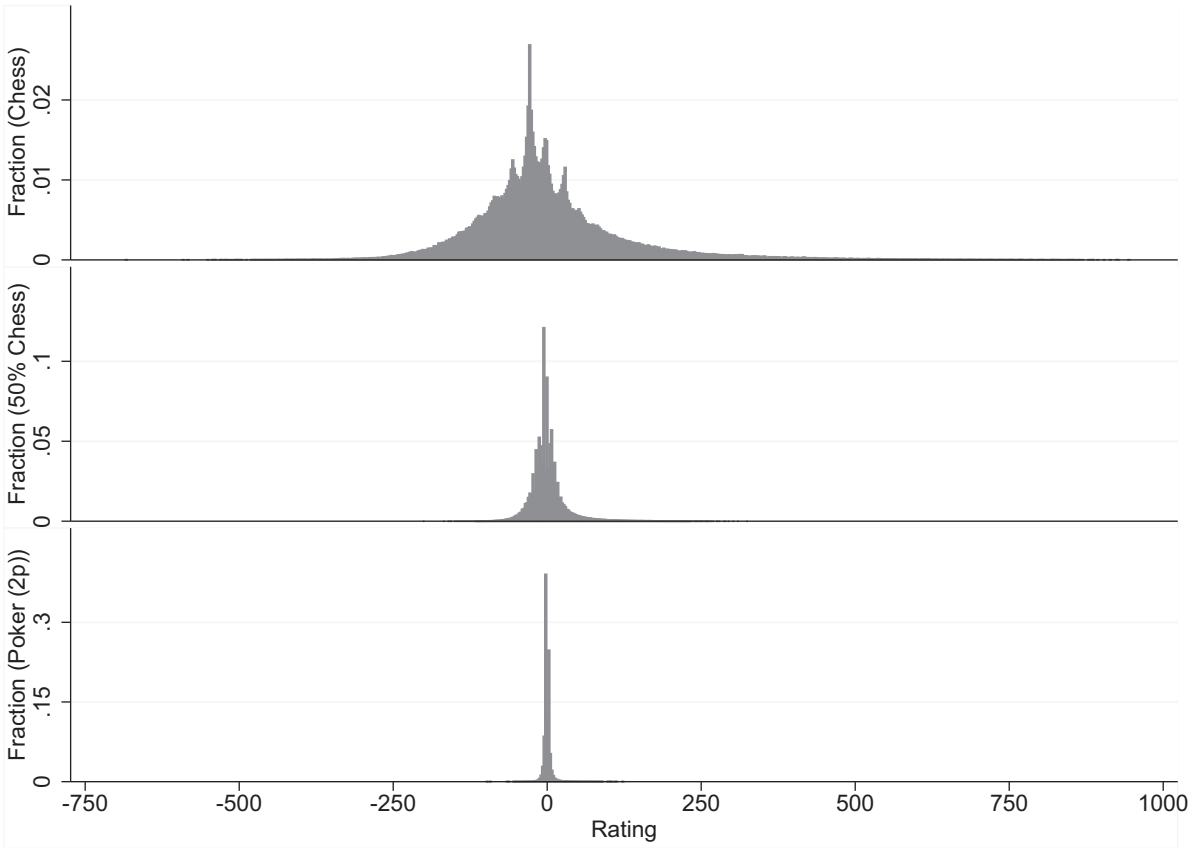

Fig. 3. Rating distributions for chess, 50%-chess, and poker (2p) - all players

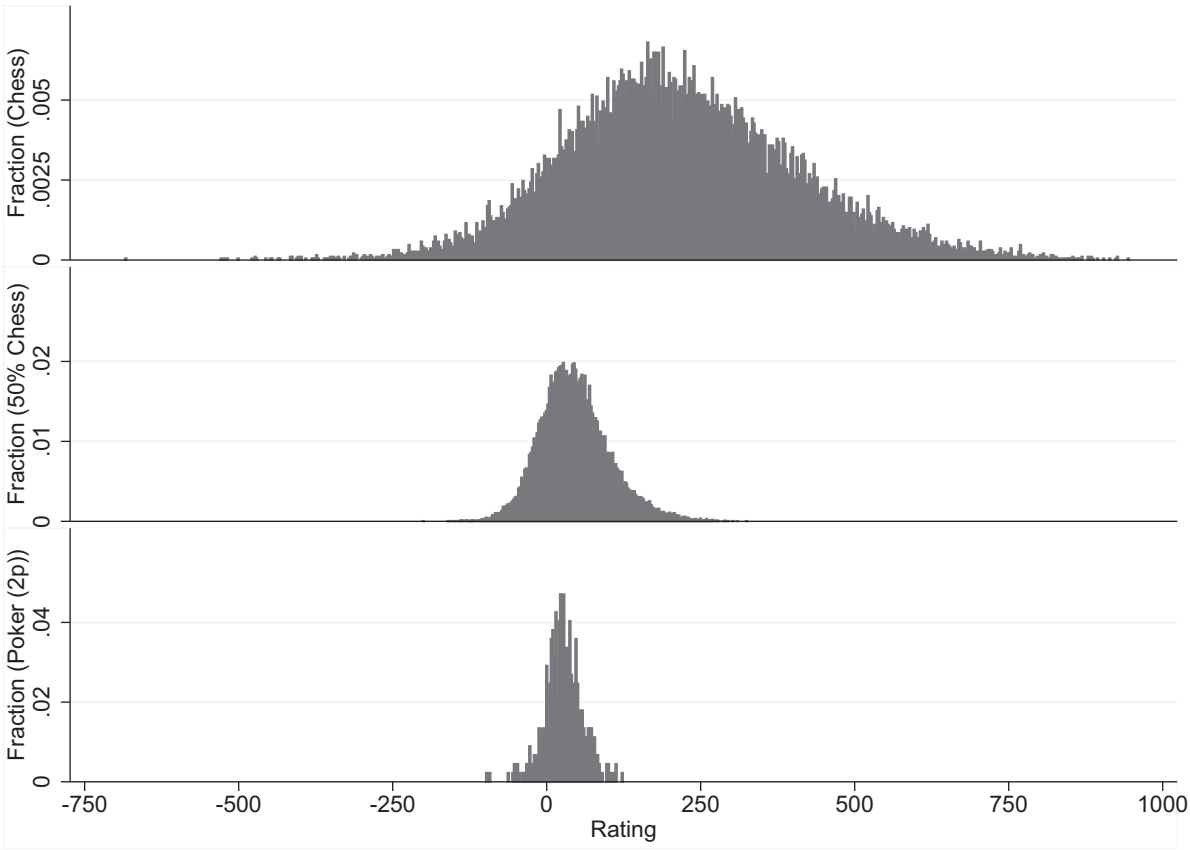

Fig. 4. Rating distributions for chess, 50%-chess, and poker (2p) - players with 100 or more games only

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.[37]

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

A3. Normalization of Elo rankings

In this subsection, we provide a formal proof of the fact that the normalization parameter in the USCF version of the Elo rating does not change our results. The main reason is our calibration of the k-factor, as the minimization process adjusts the optimal k-factor accordingly. Thus, the results are equal apart from scaling. The USCF uses

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

to calculate expected outcomes. Furthermore, the update formula to adjust ratings after each observation of S_{ij}^t is

$$R_i^{t+1} = R_i^t + k \cdot \left(S_{ij}^t - E_{ij}^t\right). \tag{A.2}$$

We show that, given a constant set of observations S_{ij}^t and equal initial ratings $R_i^0 = \hat{R}_i^0$, the expectation formula

$$\hat{E}_{ij}^t := \frac{1}{1+10^{-(\hat{R}_i^t - \hat{R}_j^t)}} \tag{A.3}$$

and the update formula

$$\hat{R}_i^{t+1} = \hat{R}_i^t + \hat{k} \cdot \left(S_{ij}^t - \hat{E}_{ij}^t\right), \tag{A.4}$$

[37] 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.

where

$$\hat{k} = \frac{1}{400} \cdot k, \tag{A.5}$$

lead to the same predictions for each and every game. The resulting ratings (and therefore, the standard deviation of their distribution) are equal apart from scaling.

Definition A.1. Two Elo ratings are ***equivalent*** if they lead to the same expected outcomes for every player and every match.

Proposition 1. *Assume a constant set of observations S_{ij}^t and equal initial ratings $R_i^0 = \hat{R}_i^0 = 0$. Then, the USCF Elo rating with expectation formula (A.1) and updating formula (A.2) is equivalent to the Elo rating with expectation formula (A.3) and updating formula (A.4).*

Proof. First, we show that

$$\hat{R}_i^t = \frac{1}{400} \cdot R_i^t \qquad \forall i, t \tag{A.6}$$

by induction over t. At $t = 0$, all ratings equal zero. Therefore, it is left to show that, given (A.6),

$$\hat{R}_i^{t+1} = \frac{1}{400} \cdot R_i^{t+1}.$$

Note that (A.6) yields

$$-(\hat{R}_i^t - \hat{R}_j^t) = -\frac{(R_i^t - R_j^t)}{400}. \tag{A.7}$$

Now, by definition,

$$\hat{R}_i^{t+1} = \hat{R}_i^t + \hat{k} \cdot \left(S_{ij}^t - \hat{E}_{ij}^t\right) = \hat{R}_i^t + \hat{k} \cdot \left(S_{ij}^t - \frac{1}{1 + 10^{-\left(\hat{R}_i^t - \hat{R}_j^t\right)}}\right).$$

Using (A.5), (A.6), and (A.7) yields

$$\hat{R}_i^{t+1} = \frac{1}{400} \cdot R_i^t + \frac{1}{400} \cdot k \cdot \left(S_{ij}^t - \frac{1}{1 + 10^{-\frac{\left(R_i^t - R_j^t\right)}{400}}}\right) = \frac{1}{400} \cdot \left(R_i^t + k \cdot \left(S_{ij}^t - E_{ij}^t\right)\right) = \frac{1}{400} \cdot R_i^{t+1}.$$

Finally, given (A.7) holds, it follows that

$$E_{ij}^t = \hat{E}_{ij}^t \qquad \forall i, j \in \{1, \ldots k\}, t \in \{1, \ldots, T\}.$$

□

Remark 1. If k^* minimizes the loss function of USCF Elo rating with expectation formula (A.1) and updating formula (A.2), then $\hat{k}^* = \frac{1}{400} \cdot k^*$ minimizes the loss function of the Elo rating with expectation formula (A.3) and updating formula (A.4).

A4. Minimization of loss function

Here we describe the numerical procedure used to minimize the quadratic loss function given in (2). 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. Our algorithm is based on 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 5, 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, 40, 80, 120 and 160.[38] Suppose 40 produces the lowest loss among those five, then we continue by halving the grid size taking 40 as center point k^*, i.e., the new grid will consist of 0, 20, 40, 60, and 80. 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 6).

Table 7 shows the results of the procedure for each dataset. It includes the optimal k-factor derived through the numerical algorithm, k^*, as well as the resulting value of the loss function $\mathcal{L}(k^*)$ when applying this k-factor to the data. The value

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

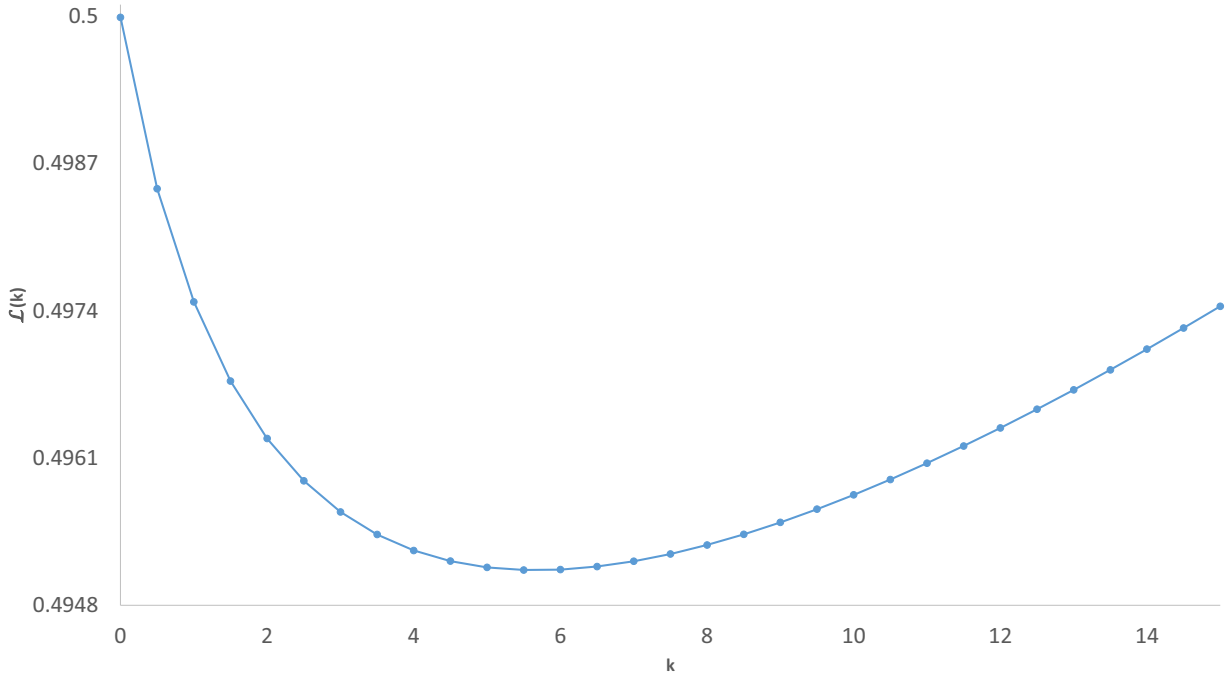

Fig. 5. Loss as function of k-factor for the game "Backgammon"

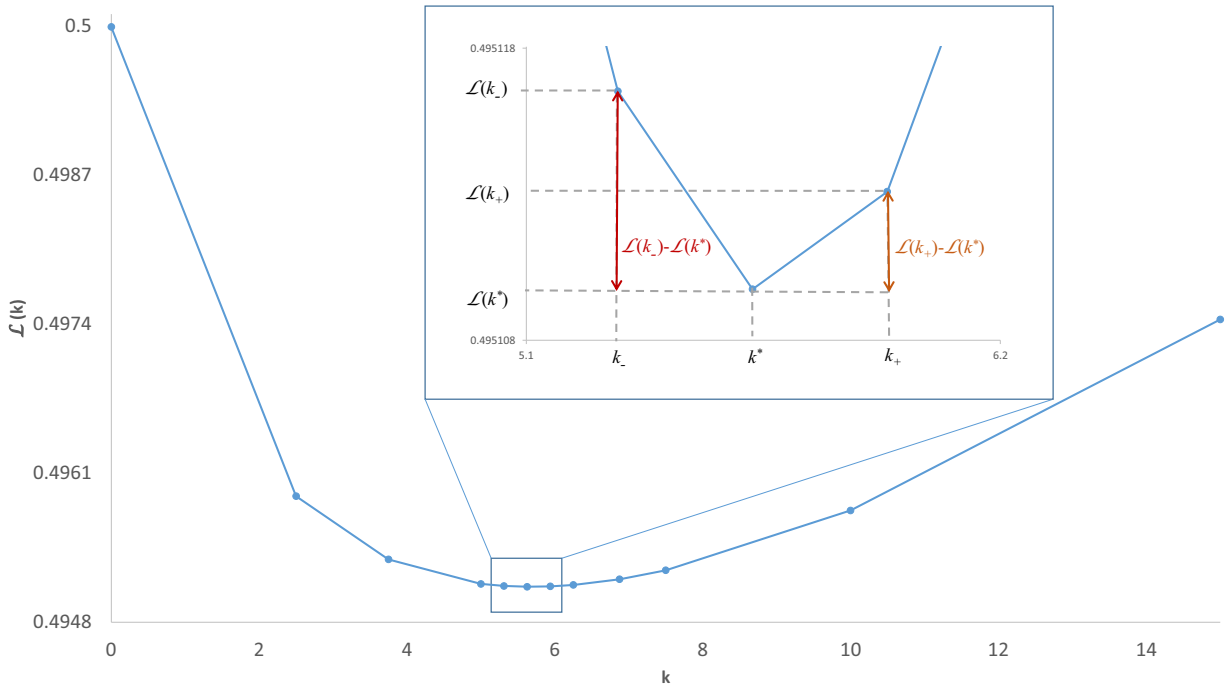

Fig. 6. Numerical procedure for the game "Backgammon"

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.[39]

[39] For games where every match has an outcome of 0 or 1, per definition $\mathcal{L}(0) = 0.5$, as this loss would result when predicting both players to be equally likely to win in each of the matches. Note that chess and 50%-chess are the only datasets which include draws. This significantly reduces the values of their loss functions, because a draw is very close to the predicted outcome whenever two players of similar skill compete with each other.

Table 7
Derived k-factors and corresponding loss-function values

	$\mathcal{L}(0)$	$\mathcal{L}(k^*)$	k^*
Go	0.5	0.384	104.1
Tennis	0.5	0.411	48.1
Chess	0.359	0.298	57.0
Tetris (2p)	0.5	0.475	39.5
Jewels (2p)	0.5	0.491	12.4
50% Chess	0.359	0.350	12.0
Rummy (2p)	0.5	0.491	9.8
Backgammon	0.5	0.495	5.7
Solitaire (2p)	0.5	0.497	4.9
Poker (2p)	0.5	0.494	4.9
Yahtzee (2p)	0.5	0.496	4.4
Crazy 8s (2p)	0.5	0.498	3.6

Table 8
Standard deviations for artificial pure chance datasets

	"Pure Chance" (Backgammon Data)	*"Pure Chance" (Poker (2p) Data)*	*"Pure Chance" (Solitaire (2p) Data)*
#Players	4,229	55,158	33,762
#Regulars	780	1,883	9,374
#Matches	42,126	191,704	641,220
Std. Dev. (All)	<0.3	<0.004	<0.003
Std. Dev. (Regulars)	<0.6	<0.015	<0.005

Table 9
Summary statistics of rating distributions - simulated x%-deterministic

	Std. Dev.	Obs.	Min.	1%	99%	Max.	p^{sd}	p_1^{99}
50% Determ.	122.8	1,000	-304.7	-263.0	244.1	295.7	67.0	94.9
40% Determ.	91.9	1,000	-232.4	-193.8	192.6	223.8	62.9	90.2
30% Determ.	61.1	1,000	-156.3	-127.6	129.4	172.1	58.7	81.4
20% Determ.	32.3	1,000	-95.9	-67.3	69.2	87.1	54.6	68.7
15% Determ.	22.7	1,000	-66.0	-51.3	50.7	65.0	53.3	64.3
10% Determ.	10.7	1,000	-32.4	-23.1	23.8	34.3	51.5	56.7

A5. Pure chance simulations

In order to interpret the results of our algorithm and compare different datasets, we want to verify whether pure chance simulations produce standard deviations close to zero like the theory would predict. Furthermore, we want to test whether the size of the dataset has any impact. For this reason, we chose three datasets of different size (backgammon, poker (2p) and solitaire (2p)) and replace the result of every match by a coinflip. Afterwards, we measure these artificial datasets with our procedure.[40] Table 8 summarizes the results.

The standard deviations of estimated playing strengths in our pure chance simulations turn out to be very small, while an increase in observations seems to drive the results even closer to the theoretical prediction of zero. The reason for this is our calibration method of the optimal k-factor. One could think of the k-factor to measure the predictive power of the observed (randomly generated) results, which by the law of large number tends toward zero if the number of observations increases.

A6. x% simulations

In Table 9, we summarize the results of our x%-deterministic simulations. We can use these results to estimate how much chance needs to be injected into the deterministic game to end up with a distribution similar to poker. The standard deviations of our poker data range from 18 to 23, while 15% deterministic has a standard deviation of 22.7. Therefore, we call poker to be roughly 15% deterministic.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.euroecorev.2020.103472

[40] Due to the computational effort, we restrict ourselves to report upper bounds instead of calculating the standard deviations to full precision.

References

Beggs, S., Cardell, S., Hausman, J., 1981. Assessing the potential demand for electric cars Journal of Econometrics 17, 1–19.
Bewersdorff, J., 2004. Luck, logic, and white lies: the mathematics of games. CRC Press.
Biais, B., Hilton, D., Mazurier, K., Pouget, S., 2005. Judgemental overconfidence, self-monitoring, and trading performance in an experimental financial market. The Review of Economic Studies 72 (2), 287–312.
Binde, P., Romild, U., Volberg, R.A., 2017. Forms of gambling, gambling involvement and problem gambling: Evidence from a Swedish population survey. International Gambling Studies 17 (3), 490–507.
Borm, P., van der Genugten, B., 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., Hannum, R., 2005. Poker: Public policy, law, mathematics, and the future of an American tradition. TM Cooley L. Rev. 22, 443.
Camerer, C., Lovallo, D., 1999. Overconfidence and excess entry: An experimental approach. American Economic Review 89 (1), 306–318.
Croson, R., Fishman, P., Pope, D.G., 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., Detterman, D.K., 2008. Poker is a skill. Gaming Law Review 12 (1), 31–36.
Dreef, M., Borm, P., van der Genugten, B., 2003. On strategy and relative skill in poker. International Game Theory Review 5 (02), 83–103.
Dreef, M., Borm, P., van der Genugten, B., 2004. Measuring skill in games: Several approaches discussed. Mathematical Methods of Operations Research 59.3, 375–391.
Dreef, M., Borm, P., van der Genugten, B., 2004. A new relative skill measure for games with chance elements. Managerial and Decision Economics 25 (5), 255–264.
Dufwenberg, M., Sundaram, R., Butler, D.J., 2010. Epiphany in the game of 21. Journal of Economic Behavior & Organization 75 (2), 132–143.
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., Rock, J.-P., 2009. Quantifying skill in games - theory and empirical evidence for poker. Gaming Law Review and Economics 13 (1), 50–57.
Filippin, A., Schmidt, U., Tomasuolo, M., 2020. Is gambling more dangerous than betting? Unpublished manuscript.
van der Genugten, B., Borm, P., 2016. Texas hold'em: A game of skill. International Game Theory Review 18 (03), 1650005.
Glickman, M.E., Doan, T., 2017. The US chess rating system. US Chess Federation.
Holznagel, B., 2008. Poker - Glücks- oder Geschicklichkeitsspiel. Multimedia und Recht 11 (7), 439–444.
Kelly, J.M., Dhar, Z., Verbiest, T., 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., Kadane, J.B., Austin, R., Zamir, S., 1997. Skill in games. Management Science 43 (5), 596–609.
Levitt, S.D., Miles, T.J., 2014. The role of skill versus luck in poker: Evidence from the world series of poker. Journal of Sports Economics 15 (1), 31–44.
Malmendier, U., Tate, G., 2005. CEO overconfidence and corporate investment. The Journal of Finance 60 (6), 2661–2700.
Malmendier, U., Tate, G., 2008. Who makes acquisitions? CEO overconfidence and the market's reaction. Journal of Financial Economics 89 (1), 20–43.
Park, Y.J., Santos-Pinto, L., 2010. Overconfidence in tournaments: Evidence from the field. Theory and Decision 69 (1), 143–166.
Potter van Loon, R.J.D., van den Assem, M.J., van Dolder, D., 2015. Beyond chance? The persistence of performance in online poker. PLoS one 10 (3), e0115479.
Rose, I.N., 2011. Poker's Black Friday. Gaming Law Review and Economics 15 (6), 327–331.
Siler, K., 2010. Social and psychological challenges of poker. Journal of Gambling Studies 26 (3), 401–420.
Van Essen, M., Wooders, J., 2015. Blind stealing: Experience and expertise in a mixed-strategy poker experiment. Games and Economic Behavior 91, 186–206.