Essays on Empirical Game Theory

Dissertation

zur Erlangung des akademischen Grades

Doctor Rerum Politicarum

An der

Fakultät für Wirtschafts- und Sozialwissenschaften

der Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Marco Lambrecht

Heidelberg, Juli 2020

Acknowledgements

During my studies and research at the University of Heidelberg, I met and worked with many people, to whom I am very grateful. Without their support this dissertation would not have been possible. They contributed towards this thesis and inspired me in many different ways. I would like to take this opportunity to individually thank everyone.

First of all, I am immensely grateful to my supervisor Jörg Oechssler, for giving me the chance to write this thesis and constantly supporting and encouraging me during the process. He always found time to help me, provide perceptive feedback and suggestions as well as moral support. Our discussions and his deep knowledge greatly improved the quality of my research projects. I am also very grateful to Christoph Vanberg, who kindly agreed to be my second supervisor. We had very productive discussions on a regular basis and an excellent working atmosphere.

Furthermore, I would like to thank Jürgen Eichberger, who provided valuable advice and especially encouraged and supported me during the final stage of my dissertation. I am also very thankful to Peter Duersch, with whom I have had the opportunity to collaborate. Working with him was a very positive and valuable experience, which helped me to become a better researcher. I would like to give special thanks to Andis Sofianos, who co-authored a project of this thesis and with whom I had a great time, not only when we worked together. I would like to extent my thanks to Stefan Trautmann and my other co-author Yilong Xu. We had a very friendly and productive collaboration.

I am indebted to Andreas Reischmann for his support, specifically during the first steps of my thesis, as well as Christian König and Christoph Becker for providing important assistance with programming and computations throughout.

My thanks extent to my past and current colleagues from the institute, Christoph Brunner, Lorenz Hartmann, Florian Kauffeldt, Christopher Kops, Isabel Marcin, Illia Pasichnichenko, David Piazzolo, Gert Pönitzsch, Rajesh Ramachandran, Hannes Rau, Andrea Sáenz de Viteri Vázquez, Robert Schmidt, Arjun Sengupta, Marek Steinke, and Martin Vollmann. We had stimulating discussions, a very friendly working atmosphere and a lot of fun together. I am grateful to Gerda Asmus, Angelika Budjan, Sven Kunze, Onno Kleen, David Vespermann and Min Xie, with whom I spent time during graduate school. Furthermore, Raghul Venkatesh and Moritz von Rohrscheidt provided valueable input that significantly improved this thesis, which I highly appreciate.

I would also like to thank Ulrike Arnold, Barbara Neef, Gabi Rauscher and Freya Schadt, whose efforts made my time at the institute both productive and very pleasurable. Financial support from the University of Heidelberg and DFG grant OE 198/5-1 for conducting experiments and presenting my work at conferences is gratefully acknowledged.

Finally and above all, I would like to thank my family for their support. Without their encouragement, this thesis would not exist.

Contents

Abstract

This thesis is a collection of studies about ratings and rankings and a study about pricing of cryptocurrencies.

In the first chapter, jointly with Peter Duersch and Jörg Oechssler we develop a universal measure of skill and chance in games. 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. 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.

The second chapter aims to measure skill and chance in different versions of online poker, using the best-fit Elo algorithm established in the first chapter. While *Texas Hold'em* arguably is the most popular version being played, the amount of skill involved might differ from other versions like *Omaha Hold'em*. Many platforms offer faster procedures to play (e.g. *"hyper turbo"*), as well as different levels of stakes. Given the richness of online poker data, it is possible to isolate the impact of these variations individually.

The heterogeneity of best-fit Elo ratings decreases in quicker competitions or with higher stakes. Meanwhile, Omaha seems to contain more elements of skill than Texas Hold'em, as its analysis shows a wider distribution of skill levels of players.

The third chapter motivates the introduction of the notion of *Independece of Alternatives* (IoA) in the context of ranking models. IoA postulates a property of independence which seems intuitively reasonable but does not exclusively hold in models where Luce's Choice Axiom applies. Assuming IoA, expected ranks in the ranking of multiple alternatives can be determined from pairwise comparisons. The result can significantly simplify the calculation of expected ranks in practice and potentially facilitate analytic methods that build on more general approaches to model the ranking of multiple alternatives.

The fourth chapter describes an experimental study on cryptocurrency markets. Jointly with Andis Sofianos and Yilong Xu we focus on potential effects of mining on pricing. Recent years have seen an emergence of decentralized cryptocurrencies that were initially devised as a payment system, but are increasingly being recognized as investment instruments. The price trajectories of cryptocurrencies have raised questions among economists and policy makers, especially since such markets can have spillover effects on the real economy. We focus on two key properties of cryptocurrencies that may contribute to their pricing. In a controlled lab setting, we test whether pricing is influenced by costly mining, as well as entry barriers to mining technology. Our mining design resembles the proof-of-work algorithm employed by the majority of cryptocurrencies, such as bitcoin. In our second treatment, half of the traders have access to the mining technology, while the other half can only participate in the market. This is designed to resemble the high entry cost to initiate cryptocurrency mining. In the absence of mining, no bubbles or crashes occur. When costly mining is introduced, assets are traded at prices more than 300% higher than fundamental value and the bubble peaks relatively late in the trading periods. When only half of the traders can mine, prices surge much earlier and reach values of more than 400% higher than fundamental value at the peak of the market. Overall, the proof-of-work algorithm seems to fuel overpricing, which in conjunction with entry barriers to mining is intensified.

Chapter 1

Measuring skill and chance in games

1.1 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, courts and lawmakers had to draw a line and often classify gambling as referring to games that "predominantly depend on chance".[1] But how can one measure whether the outcome

[1]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

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

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

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 skill (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.[2] 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. (2008a) 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).

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. (2003,0,0), 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

[2]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).

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 3.2 we explain our new approach for measuring skill and chance in detail. Section 2.2 describes our data and in Section 1.4 we present the empirical results. Section 1.5 concludes.

1.2 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 1.2.3 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.[3] 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.[4] In a game of pure skill like chess, the standard deviation is very high.[5]

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.[6]

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

[4] We confirm this by replacing all outcomes by random results in three datasets. The standard deviations take values of between .3 for the smallest to 0.003 for the largest dataset. For details, see Section 1.6.5.

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

[6] 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.

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.[7]

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.[8] 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.

1.2.1 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 $i, j \in I$.[9] 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. 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

$$E_{ij}^t := \frac{1}{1 + 10^{-\frac{R_i^t - R_j^t}{400}}}. \tag{1.1}$$

Expected scores range from zero (sure loss) to one (sure win). The parameter 400 in the

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

[8] cf. footnote 1.

[9] Matches are ordered chronologically by start time.

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

logit function is a normalization used by chess federations which we retain for familiarity.[11] Given this parameter, a rating difference of 100 translates into an expected score of .64. We normalize the initial rating of each player to $R_i^0 = 0$.[12] 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. $\forall i \notin \rho(t)\colon R_i^{t+1} = R_i^t$.

1.2.2 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 $t \in T$ 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

[11] Due to our calibration method, the use of this parameter is without loss of generality (Appendix 1.6.3 shows this formally).

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

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

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

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

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.

$$P^t(q^t(i) = k) = \sum_{q^t \in \{q_k^t = i\}} \prod_{l=1}^{n^t-1} \frac{e^{R_l^t}}{\sum_{j=l}^{n^t} e^{R_j^t}}$$

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

$$\hat{S}_i^t = \frac{1}{\pi_{\max}^t} \cdot S_i^t, \qquad \hat{E}_i^t = \frac{1}{\pi_{\max}^t} \cdot E_i^t.$$

Due to the normalization, these shares range from zero to one.[13]

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

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

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

Before we measure the standard deviation, we multiply all ratings by 400 to retain com-

[13]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.

parability with our two player ratings and established chess ratings.

1.2.3 Calibrating the Elo ratings

While the actual scores S_i^t are observed in our data, the expected scores are determined recursively and depend on k. To indicate this we write $E_i^t(k)$. 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 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.[14]

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:

$$k^* := \arg\min_k \frac{1}{T} \sum_{\substack{t \in T \\ i \in \rho(t)}} \left(S_i^t - E_i^t(k)\right)^2 . \tag{1.2}$$

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.[15] 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.,

[14]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).

[15]See Appendix 1.6.4 for an example graph as well as an exact description of the numerical procedure.

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.

1.3 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).[16] The remaining datasets are still quite large and are summarized by the statistics in Table 1.1. For each game, Table 1.1 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.

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.[17] 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.[18] 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

[16] 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.

[17] 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.

[18] 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 3.2.

Table 1.1: 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

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.[19] 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.[20]

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.[21] 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.[22]

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

[20] For a detailed description of these games, see Appendix 1.6.2.

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

[22] 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.

1.4 Results

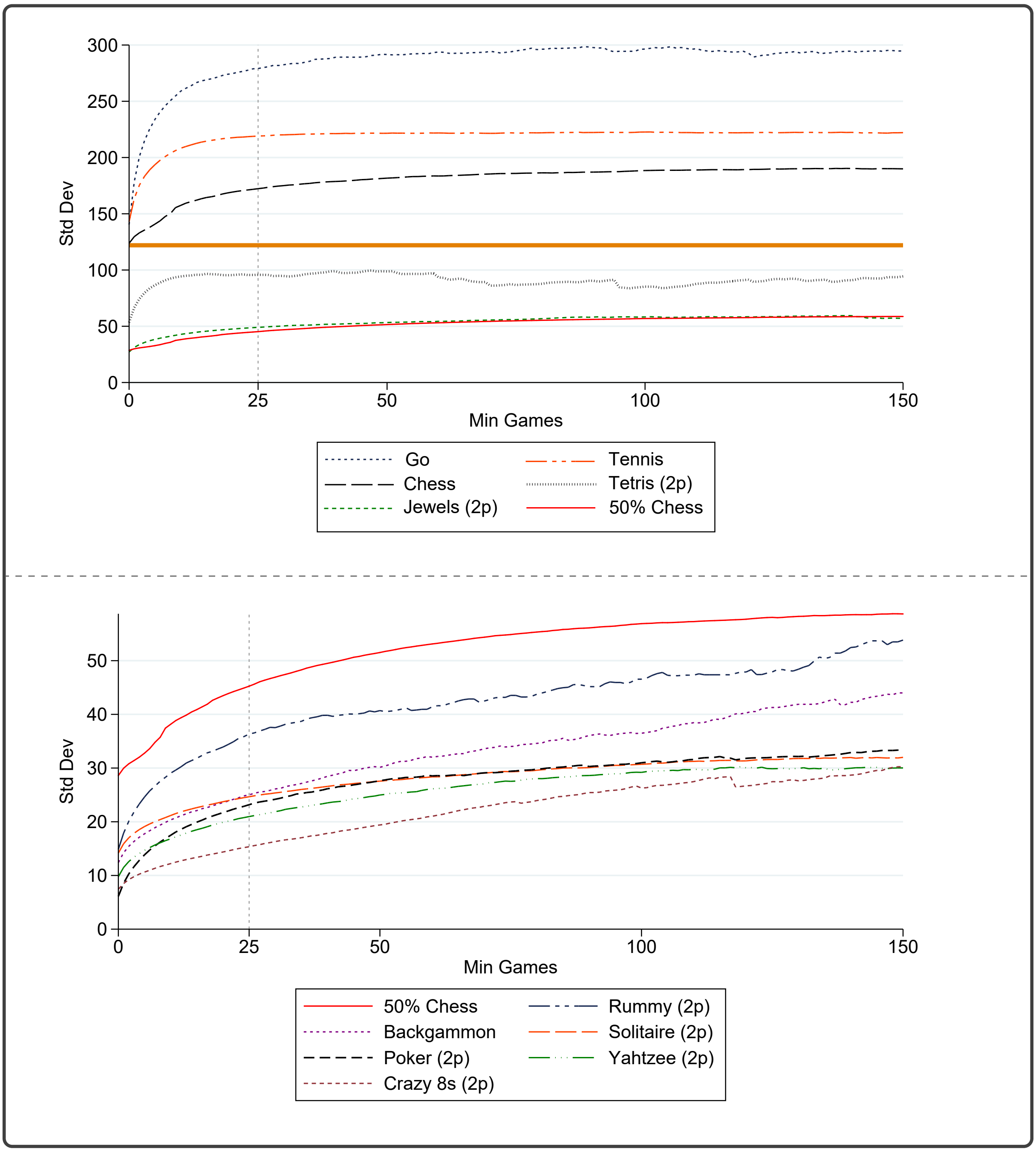
In this section, we present the results of our analysis, in particular, the standard deviations of the respective best-fit Elo rating distributions, which resulted from our rating procedure described in Section 1.2.3. Due to the fact that the Elo rating is based on an updating formula, the approximations 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 1.1 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 1.1 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 1.1. 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.[23] 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 1.6.1.

In Table 1.2 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

[23]This also holds true for the respective graphs of the multiplayer games.

Figure 1.1: 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.



99% percentile player, and most importantly, the standard deviation of all ratings. We sort the table according to this value. Via formula (1.1), 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%.[24] This number is reported in the repetitions column (abbreviated "Rep.").

Table 1.2: Summary statistics of rating distributions - regulars only

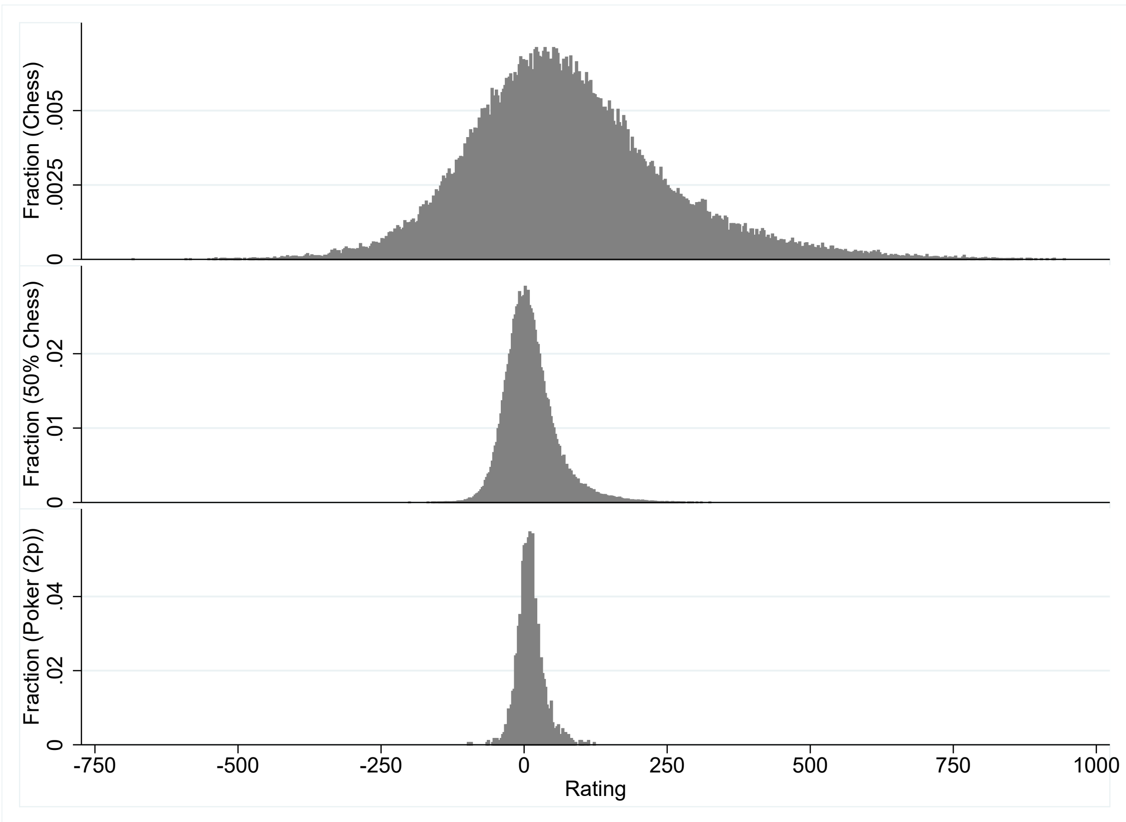
	Std. Dev.	**Min.**	**1%**	**99%**	**Max.**	**p^{sd}**	**p$_1^{99}$**	**Rep.**
Go	278.9	-659.2	-440.1	859.3	1,432.2	83.3	99.9	1
Tennis	218.8	-389.6	-242.9	790.4	1,438.0	77.9	99.7	1
Chess	171.7	-684.6	-289.4	579.6	945.3	72.9	99.3	3
50% Determ.	122.8	-304.7	-263.0	244.1	295.7	67.0	94.9	5
Tetris (2p)	95.7	-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

Table 1.2 confirms in more detail what was already apparent from Figure 1.1. 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,[25] and almost all of the other

[24]This definition is used by e.g. Potter van Loon et al. (2015).
[25]Note that 50%-chess has a standard deviation that is substantially below 50% of the standard deviation

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 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.[26]

Figure 1.2: Rating distributions for chess, 50%-chess, and poker (2p) - regulars



of chess.

[26]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.

The histograms in Figure 1.2 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.[27] 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 (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.[28] 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

[27] For details on this procedure, see Appendix 1.6.6

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

matched with opponents of similar strength.[29] 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 1.3. 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.[30]

Table 1.3: 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

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. (2008a). Whenever a player competes in a match, we use his previous results to calculate his average performance in the past.[31] 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

[29] 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.

[30] 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.

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

performance on the outcome of the current match.[32]

$$S_{ij}^{t} = \beta_0 + \beta_1 \cdot \bar{S}_{i}^{t-1} + \varepsilon_{i}^{t} \tag{1.3}$$

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.

Table 1.4: Coefficients, standard errors clustered on player level, t-values and R^2-values for regression specification (1.3)

	#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

*** $p < 0.001$

We run OLS regressions with standard errors clustered on the player level. Table 1.4 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 1.4, skill plays a statistically significant role. We can thus confirm the results of earlier studies for poker, in particular, Croson et al. (2008a). Remarkably, the coefficients in Table 1.4 have a very 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 1.4 are presented in the same order as in Table 1.2.

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

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

1.5 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.[33]

One inevitable feature of any empirical approach is that our results depend on the population we observe.[34] 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

[33] One could also adjust the method for 50%-chess and calculate expected scores according to $\hat{E}_{ij}^t = 0.5 \cdot 0.5 + 0.5 \cdot E_{ij}^t$, 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 $p_1^{99} = 72.4$, which would not change our main results.

[34] 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.

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.[35]

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 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.

1.6 Appendix

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

As a robustness check, Table 1.5 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

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

facilitate comparison, the games are presented in the same order as in Table 1.2. It is noticeable that the standard deviations of all games are substantially lower than in Table 1.2. 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 1.5: 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

Table 1.6 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 1.2. 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 1.3 and 1.4 show the histograms corresponding to Figure 1.2 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.

Table 1.6: Summary statistics of rating distributions - players with 100 or more games only

	Std. Dev.	**Obs.**	**Min.**	**1%**	**99%**	**Max.**	$\mathbf{p}^{sd}$	$\mathbf{p}_1^{99}$	**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

Figure 1.3: Rating distributions for chess, 50%-chess, and poker (2p) - all players

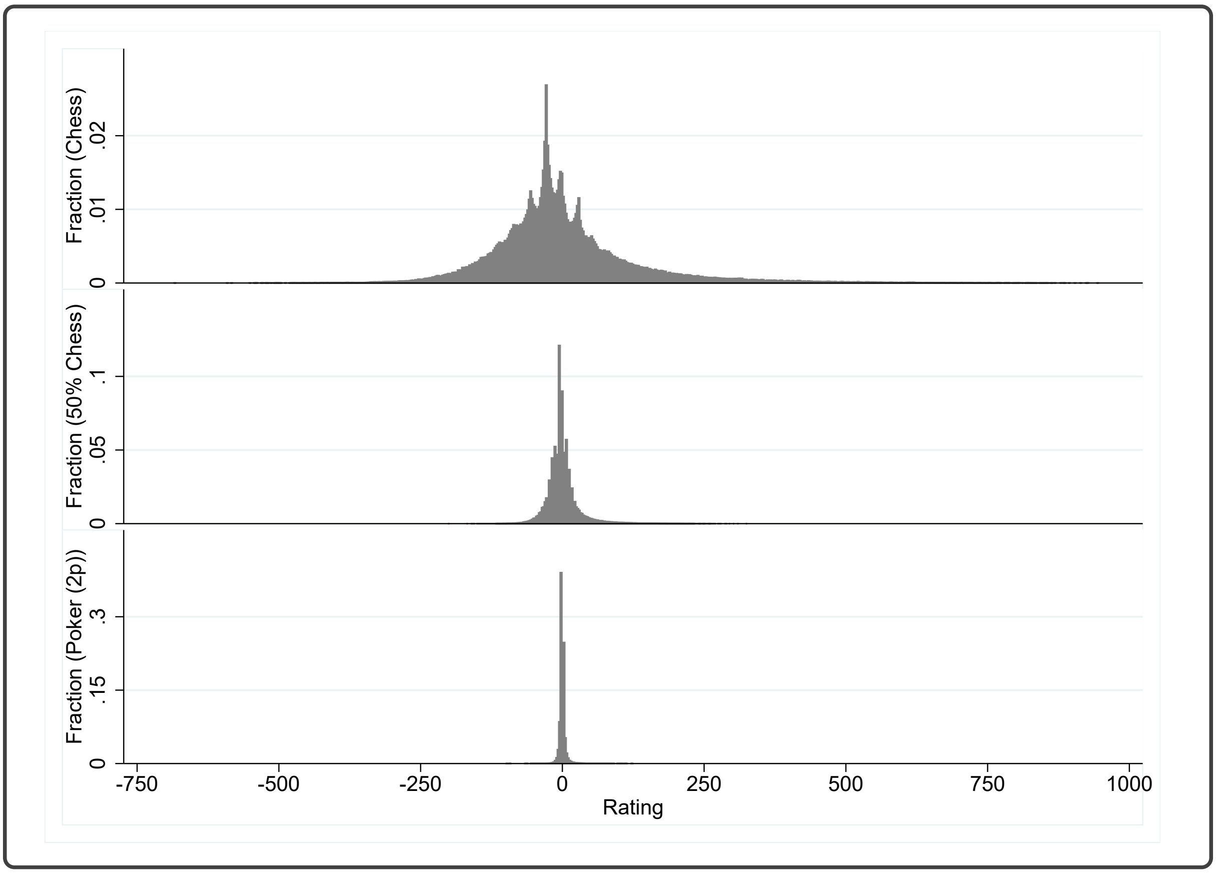
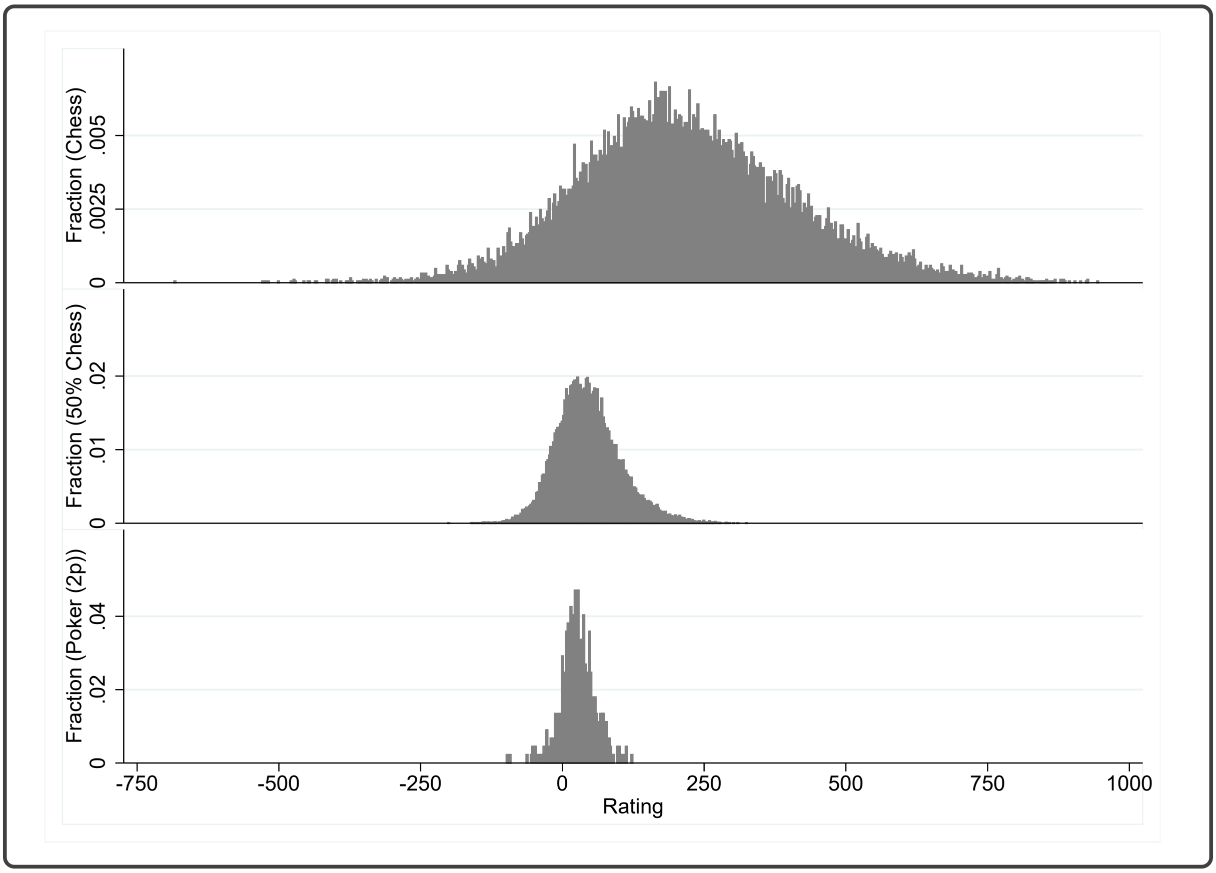


Figure 1.4: Rating distributions for chess, 50%-chess, and poker (2p) - players with 100 or more games only



1.6.2 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 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.[36]

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.

[36]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.

1.6.3 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

$$E_{ij}^t := \frac{1}{1 + 10^{-\frac{R_i^t - R_j^t}{400}}}, \tag{1.4}$$

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

$$R_i^{t+1} = R_i^t + k \cdot \left(S_{ij}^t - E_{ij}^t \right). \tag{1.5}$$

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

$$\hat{E}_{ij}^t := \frac{1}{1 + 10^{-(\hat{R}_i^t - \hat{R}_j^t)}} \tag{1.6}$$

and the update formula

$$\hat{R}_i^{t+1} = \hat{R}_i^t + \hat{k} \cdot \left(S_{ij}^t - \hat{E}_{ij}^t \right), \tag{1.7}$$

where

$$\hat{k} = \frac{1}{400} \cdot k, \tag{1.8}$$

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

Proposition 1.1. *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* (1.4) *and updating formula* (1.5) *is equivalent to the Elo rating with expectation formula* (1.6) *and updating formula* (1.7)*.*

Proof. First, we show that

$$\hat{R}_i^t = \frac{1}{400} \cdot R_i^t \qquad \forall i, t \tag{1.9}$$

by induction over t. At $t = 0$, all ratings equal zero. Therefore, it is left to show that, given (1.9),

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

Note that (1.9) yields

$$-(\hat{R}_i^t - \hat{R}_j^t) = -\frac{(R_i^t - R_j^t)}{400}. \tag{1.10}$$

Now, by definition,

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

Using (1.8), (1.9), and (1.10) yields

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

Finally, given (1.10) holds, it follows that

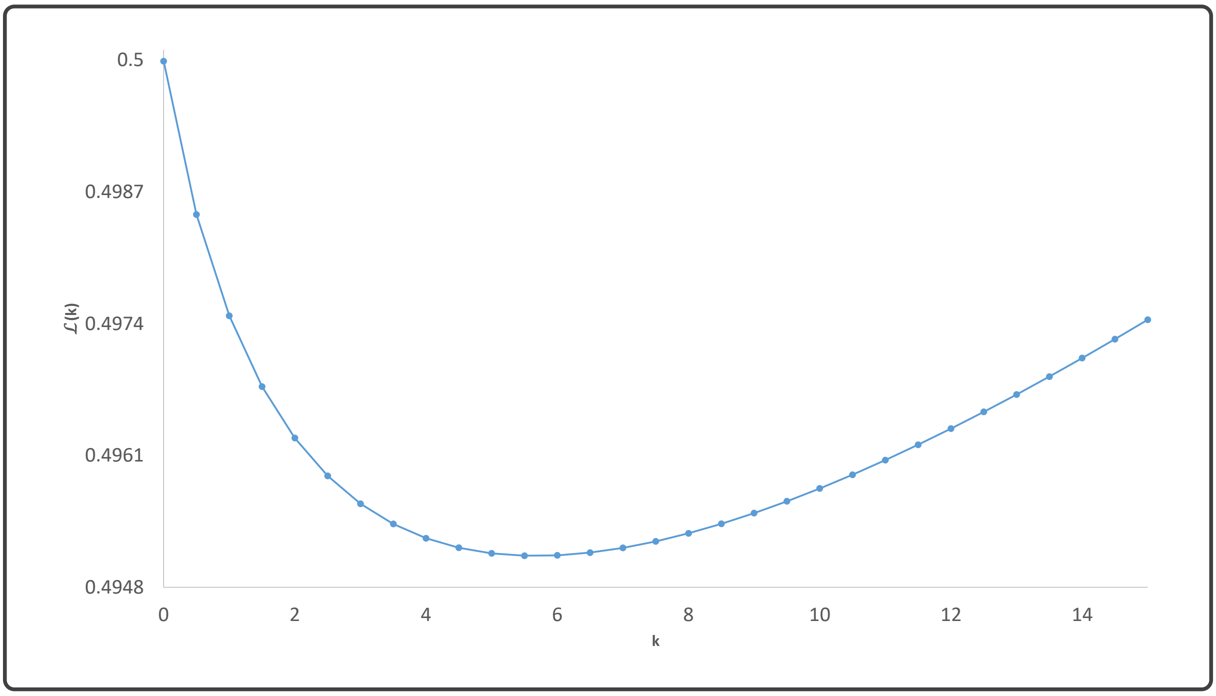
$$E_{ij}^t = \hat{E}_{ij}^t \qquad \forall i,j \in \{1,...k\}, t \in \{1,...,T\}.$$

□

Remark 1.1. *If k^* minimizes the loss function of USCF Elo rating with expectation formula* (1.4) *and updating formula* (1.5) *, then $\hat{k}^* = \frac{1}{400} \cdot k^*$ minimizes the loss function of the Elo rating with expectation formula* (1.6) *and updating formula* (1.7)*.*

1.6.4 Minimization of loss function

Figure 1.5: Loss as function of k-factor for the game "Backgammon"

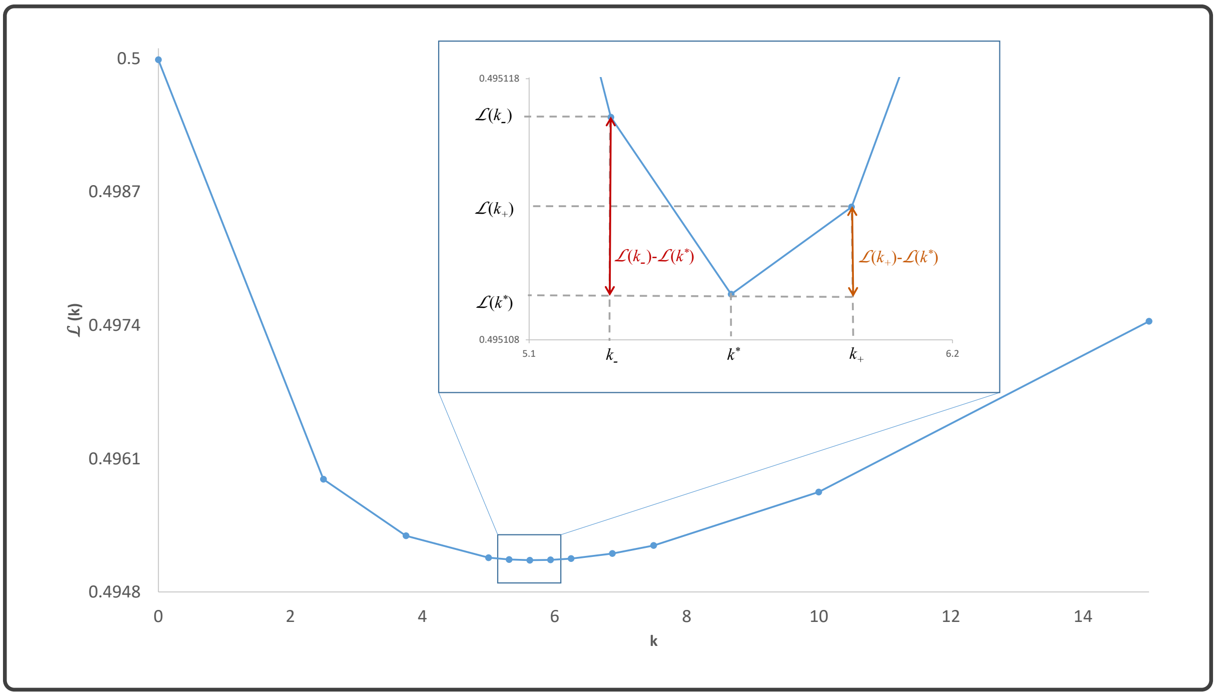


Here we describe the numerical procedure used to minimize the quadratic loss function given in (1.2). Let

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

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 1.5, which shows the loss for the game of backgammon.

Figure 1.6: Numerical procedure for the game "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.[37] 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 1.6). Table 1.7 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 of the loss function

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

Table 1.7: 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

can be interpreted similar to the Brier score (Brier, 1950). The lower this value, the more accurate are the predictions of outcomes.[38]

1.6.5 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.[39] Table 1.8 summarizes the results.

Table 1.8: 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

[38]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.

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

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.

1.6.6 x% simulations

In Table 1.9, 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.

Table 1.9: 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

Chapter 2

Measuring skill and chance in different versions of poker

2.1 Introduction

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

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

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

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

In this study, I implement the best-fit Elo algorithm (Duersch et al., 2020) to quantify the degree of skill and chance in different versions of two-player online poker. Specifically, I contrast *Texas Hold'Em* and *Omaha* poker, *regular* and *hyper turbo* settings, and stake levels of 3.50$, as well as 60$. The results show that heterogeneity of playing strengths in poker depend on the version played, speed, and stakes. In general, faster play and higher stakes decrease heterogeneity, indicating a higher influence of chance. Additionally, Omaha exhibits a higher heterogeneity of skill than No Limit Hold'em.

The next section describes poker platforms, and the data used for the analysis. Section 2.3 provides the results, and section 3.4 concludes.

2.2 Data and Procedure

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

Additionally, PokerStars offers different speeds of play by varying chip endowments in tournaments. While *regular* (REG) tournaments start with 1500 chips per player, *hyper* *turbo* (HT) tournaments endow players with 500 starting chips. As the blind structure (i.e., the enforced bets in each hand) of both tournaments is the same[40], the smaller amount of starting chips leads to less wiggle room and thus (on average) to fewer hands needed to determine the winner of the tournament. Note that HT tournaments also impose a stricter time limit on each decision.

Clearly, poker can be played with different stakes. To investigate the impact on heterogeneity of playing strength, data from both medium stake (MS) 3.50$ and high stake (HS) 60$ tournaments are included in the analysis.

Overall, this study analyzes five datasets which are summarized in Table 2.1.

Table 2.1: Poker data included in this study

	Texas	Omaha
REG	MS	MS
HT	MS/HS	MS

The datasets consist of so-called "Heads-Up Sit-and-Go"-tournaments, which will start whenever two players sit down at the same table. Note that at this point both players pay the entry fee of the tournament. Then, they are endowed with an equal amount of chips and play until one player lost all his chips to the other player. Subsequently, the winner (i.e., the player holding all the chips) will be rewarded with money worth twice his entry fee.[41]

Table 2.2 summarizes the size of the different datasets. Specifically, it describes the datasets with respect to "Regulars", which are defined as players who have played at least 25 matches.

Table 2.2: Statistics on matches, players and regulars in the different poker datasets

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

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

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

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

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

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

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

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

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

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

Here, S_{ij}^t denotes the observed score of player i in match t.[42] The ratings of players who are not involved in match t do not change. The best-fit Elo algorithm calibrates the parameter k for each dataset individually. In order to achieve the best possible calibration, the optimal value k^* is chosen as:

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

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

Once all players are rated according to the best-fit Elo algorithm, the focus is on the standard deviation of the rating distributions of each game. In the Elo rating, a given difference in ratings of two players corresponds directly to the winning probabilities when

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

the two players are matched against each other. Thus, the more heterogeneous the ratings are, the better one can predict the winner of a match. If the distribution of Elo ratings is very narrow, then even the best players are not predicted to have a winning probability much higher than 50%. The wider the distribution, the more likely are highly ranked players to win when playing against lowly ranked players, and the more heterogeneous are the player strengths. In my data, the rating distributions of all games are unimodal,[43] which makes it possible to interpret the standard deviation of ratings as a measure of skill. For further details on the best-fit Elo algorithm (including an extention to multiplayer cases), see Duersch et al. (2020).

2.3 Results

Following Duersch et al. (2020), I focus the analysis on "Regulars" (players that have played at least 25 matches).[44] Table 2.3 reports the result of the analysis. Note that the focus of measurement is the standard deviation of Elo rating distributions of regular players. The table reports the minimum and maximum rating, and the rating of the 1% and the 99% percentile player. One can transform the standard deviation of each game into the corresponding winning probability of a player who is exactly one standard deviation better than his opponent. This probability is denoted as p^{sd}. For comparison, the table also provides the winning probablities when a 99% percentile player is matched against a 1% percentile player, which is denoted as p_1^{99}. The winning probability p_1^{99} can be used to calculate the number of matches necessary so that a player who is in the top percentile wins more than half of the matches with a probability larger than 75% against an opponent that is in the bottom percentile.[45] This number is reported in the repetitions column (abbreviated "Rep.").

Overall, the heterogeneity of playing strengths in the different versions of poker is moder-

[43]See Figure 2.2 in the Appendix.
[44]Due to the updating nature of Elo ratings, initial ratings might not approximate true playing strengths well. On the other hand, setting the threshold too high might thin out the data. Figure 2.1 in the Appendix depicts the standard deviation of ratings for different thresholds of minimum games, where a threshold of 25 matches seems reasonable.
[45]This definition is used by Potter van Loon et al. (2015).

Table 2.3: Summary statistics on the distribution of Elo ratings in the different poker datasets. Note that, in contrast to chess, ratings are centered on zero by design.

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

ate. This data suggests that even the best players do not win much more than two thirds of their matches when playing against the worst players.

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

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

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

2.4 Conclusion

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

Comparing Texas Hold'em and Omaha, it turns out that Omaha contains more skill elements than Texas Hold'em. Comparing the rules of the Omaha and the Texas version, it is worth noting that Omaha players have additional private information due to the fact that they hold four private cards instead of two. This seems to increase the complexity significantly, for example when considering the calculation of winning probabilities before all community cards are revealed. Even after all cards are revealed, Texas Hold'em players have to evaluate 21 different five card combinations out of the community cards and their private cards to determine their strongest hand, while Omaha players have to consider 60 different combinations.[46]

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

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

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

Due to the differences in popularity, the size of the datasets for the different versions is not balanced. Specifically, Omaha-REG seems to be played by few players, while in general

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

HT tournaments are a lot more popular than REG tournaments. However, as Duersch et al. (2020) show, the best-fit Elo algorithm measures skill independent of the size of the dataset, conditional on a minimum amount of data to approximate well. One might argue that the smaller datasets of this study could be in question to meet that requirement.[47] It seems fair to assume though that the overall trend, namely that Omaha exhibits a larger amount of skill than Texas Hold'em, would not be reversed if more data was available.

In the context of professional poker players, it is worth noting that the profitability of playing a certain type of poker (while assuming to be a "winning player") not only depends on the winning probability, but also on the duration of play. Therefore, it is not clear whether Omaha players can potentially earn the most, as tournaments might take longer. The same argument might decrease the differences in profitability when comparing HT to REG tournaments, as one can potentially play several HT tournaments in the period of time it takes to play one REG tournament. Furthermore, variations in popularity of different versions influence the search costs to find a suitable tournament to play. In fact, given the relatively small amount of data that was available on Omaha-REG which reflects the frequency of play on the website, frictions might arise because of a limited number of opponents (and thus, opportunities to play).

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

2.5 Appendix

Figure 2.1: Standard deviation of rating distributions for different cut-off values (min. number of matches per player). The vertical dashed line indicates a minimum of 25 games.

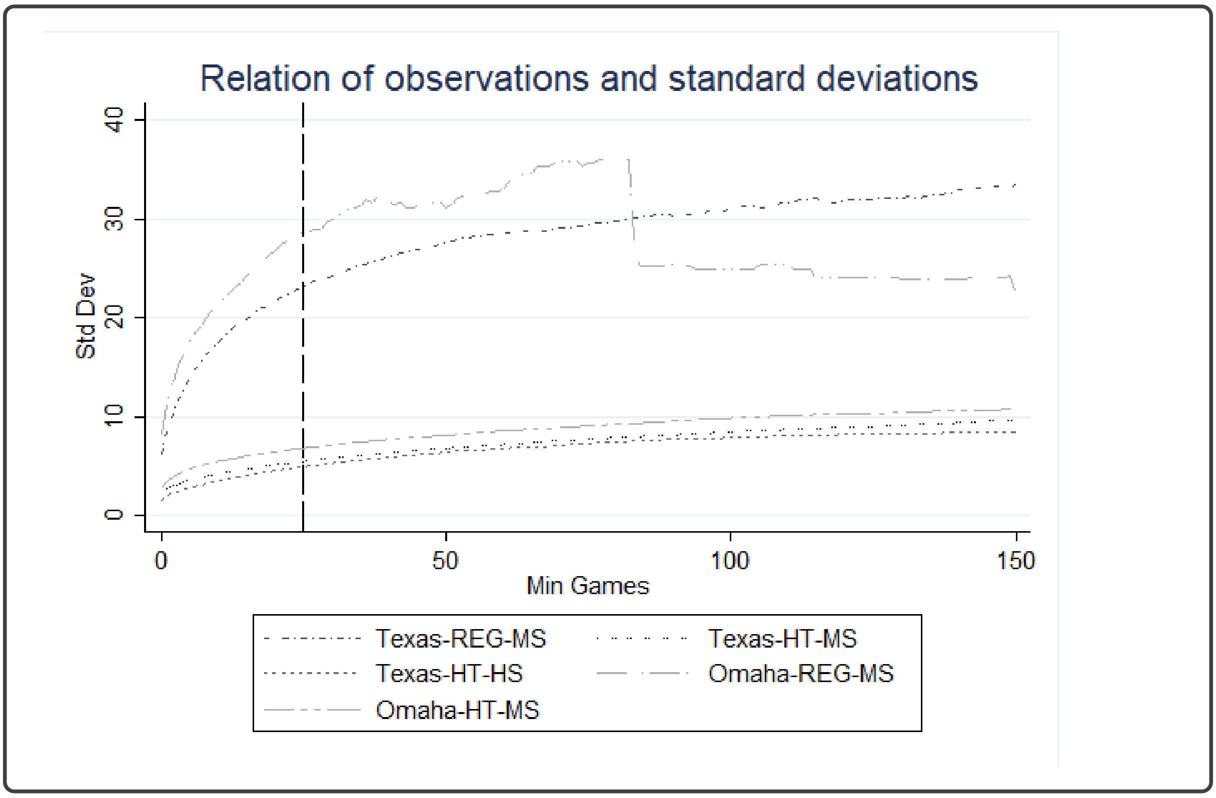
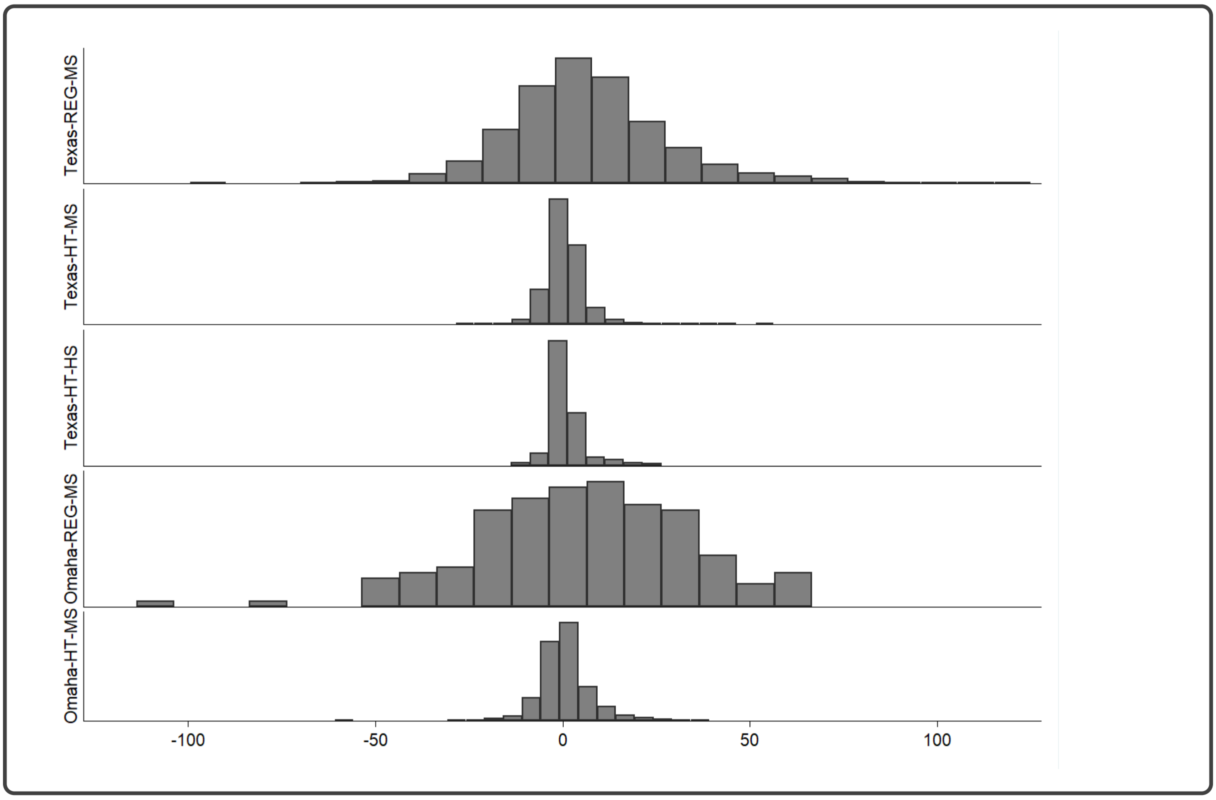


Figure 2.2: Rating distributions for different versions of poker.



Chapter 3

Independence of Alternatives in ranking models

3.1 Introduction

Luce's Choice Axiom (LCA, Luce (1959)) is one of the strong axioms in the field of ranking multiple alternatives. Breitmoser (2019) showed that, under the assumption that Positivity holds, it is equivalent to the property of Independence of Irrelevant Alternatives (IIA) introduced by Luce (1959) and restated by McFadden (1973).[48] Given the implications of LCA, which (among others) have been studied by Yellott Jr (1977) and Allison and Christakis (1994), one might wonder whether, in fact, the property of IIA could potentially be different from what its intuitive meaning suggests. On the one hand, one might wonder about the meaning of "irrelevant" in models where positivity holds. Furthermore, "independence" could potentially be interpreted differently, as Debreu's critique (Debreu, 1960) extends from LCA to IIA.

Consider the following situation for illustration. A government tries to aggregate voters' opinions in order to choose from different policies, but not all policies are available at all times. Now, one could ask whether the relative order of two policies depends on the availability of other policies. When Luce motivates IIA (Luce (1959), p. 9), he intuitively

[48]Note that Luce in his original work showed that LCA and Positivity imply IIA, but did not show equivalence.

describes it as follows:

"The idea states that if one is comparing two alternatives according to some algebraic criterion, say preference, this comparison should be unaffected by the addition of alternatives or the subtraction of old ones (different from the two under consideration)."[49]

If the government assumes that this idea is reasonable, it might infer that IIA and (given positivity) LCA hold in this situation. In the opinion of the author, this is questionable, as IIA and LCA do not capture the intution described above.

On the one hand, imagine the government would set up a simple urn model to produce the ranking of policies. Let us assume that policies are represented by colors, with blue and red balls in the urn, and the number of blue and red balls determines the chance of one color being drawn first. The order in which the colors are drawn determine their ranking. Now, we can also add additional black balls to the urn, keeping the blue and red balls constant. In this situation LCA holds, as the ratio of blue and red balls in the urn does not change. Now imagine that we change the rules slightly, introducing that the ball which is drawn first is actually ranked last, the next distinct color drawn is ranked second-to-last and so on. This obviously changes the probabilities to be ranked first (which corresponds to being drawn third in this example) and, for any case of more than two colors, LCA (as well as IIA) is violated.[50] However, it is not straightforward to see why this model should violate "independence", as we do not change the composition of the urn with respect to blue and red balls while adding black balls.

On the other hand, note that IIA only imposes conditions on probabilities to rank first. Specifically, assume that policies X and Y are among the alternatives to be ranked, and X will be ranked either first or second, if Y is in the set to be ranked. At the same time X will rank either first or last, if Y is not in the set. In the light of Luce's description, it might seem unintuitive to say that the situation described can satisfy IIA, as the relative ranking of X and a third option, say Z, seems to depend on whether Y is to be ranked or not. Yet, a situation like this is not excluded by IIA, see the following example:

[49] Luce then argues that the "probabilistic analogue of this idea is not perfectly clear", and proposes to translate it as constant ratios of the probabilities to be ranked first.

[50] Instead, a modified version holds, where the probability ratios of being ranked last are constant across different sets. This modified version was discussed by Luce himself, see Theorem 8 of Luce (1959).

Example 3.1. *Let rankings be denoted by vectors, i.e.* (A, B, C, D) *implies item* A *is ranked first,* B *second, and so on. Let* $C = \{W, X, Y, Z\}$*. Now, define the following probabilities on rankings:*

$$Pr_C((W, X, Y, Z)) = Pr_C((X, Y, W, Z)) =$$
$$Pr_C((Y, X, W, Z)) = Pr_C((Z, X, Y, W)) = 0.25$$

Note that X *is ranked either first or second, while all elements have an equal chance to end up first. In sets of three alternatives, define*

$$Pr_{C\setminus\{Z\}}((W, X, Y)) = Pr_{C\setminus\{Z\}}((X, Y, W)) = Pr_{C\setminus\{Z\}}((Y, X, W)) = \frac{1}{3}$$
$$Pr_{C\setminus\{Y\}}((W, Z, X)) = Pr_{C\setminus\{Y\}}((X, Z, W)) = Pr_{C\setminus\{Y\}}((Z, W, X)) = \frac{1}{3}$$
$$Pr_{C\setminus\{X\}}((W, Z, Y)) = Pr_{C\setminus\{X\}}((Z, Y, W)) = Pr_{C\setminus\{X\}}((Y, Z, W)) = \frac{1}{3}$$
$$Pr_{C\setminus\{W\}}((Z, X, Y)) = Pr_{C\setminus\{W\}}((X, Y, Z)) = Pr_{C\setminus\{W\}}((Y, X, Z)) = \frac{1}{3}$$

Again, all elements have an equal chance to end up first. However, X *is exclusively ranked first or second, if* Y *is in the set. In the situation where* Y *is missing,* X *ends up either first or last. Finally, assume that in all pairwise comparisons, all items have equal probabilities of 0.5.*

It is straightforward to see that probabilities to rank first are equal for all alternatives in any subset of C (i.e., constant probability ratios of exactly one). Therefore, IIA as introduced by Luce formally holds, while the situation might violate what Luce expresses to be his intuition of IIA.

In this paper, I propose the property of "Indepence of Alternatives" (IoA). It attempts to capture Luce's intuition, while providing structure that can potentially be beneficial in determining expected ranks of alternatives. Imagine the government wants to aggregate the rankings of available policies applying the "Rule of Borda". Assuming IoA, it can take a shortcut in ranking policies. Instead of eliciting full rankings of every voter and applying the rule, it is sufficient to know the proportion of voters that favors one policy over another in pairwise comparisons.

Note that other authors have promoted different notions of independence in the context of stochastic choice. Yet, IoA differs from *"i-independence"* introduced by Manzini and Mariotti (2014) and *"independence"* introduced by Gul et al. (2014). In this comparison, one might say that IoA is an axiom in the context of stochastic *ranking* rather than stochastic *choice*, as it does not (directly) impose restrictions on probabilities to be chosen, but to be ranked better. While stochastic choice and IIA helps to answer the question "Which alternative is the best?", stochastic ranking rather addresses the question "Which alternative is better?".

Given that IoA holds, I present a result to determine the expected rank of an item in the ranking of n alternatives from only pairwise comparisons. It relates to results by Luce (1959) and McFadden (1973), who established a similar link for LCA and IIA models, namely that the probability to be ranked first in any set can be inferred from pairwise comparisons. The differences of these results might become apparent when considering them in the context of chess ratings, see Chapter 3.3 for details.

The next chapter introduces mathematical properties associated with the ranking of n alternatives and describes their mutual relations. Chapter 3.3 derives the result for expected ranks and discusses its implications, and Chapter 3.4 concludes.

3.2 Mathematical properties of rankings

3.2.1 Notation

This paper studies rankings of multiple alternatives. Note that ties are excluded in these rankings. To fix notation, let

- C be a finite set of n objects,
- Σ be the power set of C,
- $(\Omega, \mathcal{A}, \mathbb{P})$ be a general abstract probability space.

Now, consider a family of rankings $\{r_S\}_{S\in\Sigma}$ on subsets of C (as $S \in \Sigma$). Rankings are random variables on Ω:

$$r_S : S \times \Omega \to \mathbb{N}$$

for all $S \in \Sigma$.

Example 3.2. *Define* $\Omega = \{\omega_1, \omega_2\}$ *and* $C = \{x, y, z\}$*. Now, define the rankings according to Table 3.1.*

Table 3.1: Ranking example

$r_C(c, \omega)$

	x	y	z
ω_1	1	2	3
ω_2	3	1	2

$r_{\{x,y\}}(c, \omega)$

	x	y
ω_1	1	2
ω_2	2	1

$r_{\{x,z\}}(c, \omega)$

	x	z
ω_1	2	1
ω_2	2	1

$r_{\{y,z\}}(c, \omega)$

	y	z
ω_1	1	2
ω_2	2	1

In this example, $r_C(z, \omega) > 1 \;\; \forall \;\; \omega \in \Omega$*, i.e. alternative* z *is never ranked first when ranking the set* C *(but it will always be ranked first when ranking* $\{x, z\}$*).*

Note that this notation allows to formulate the probability for item $c \in C$ to be ranked at position i when ranking the set S, i.e.

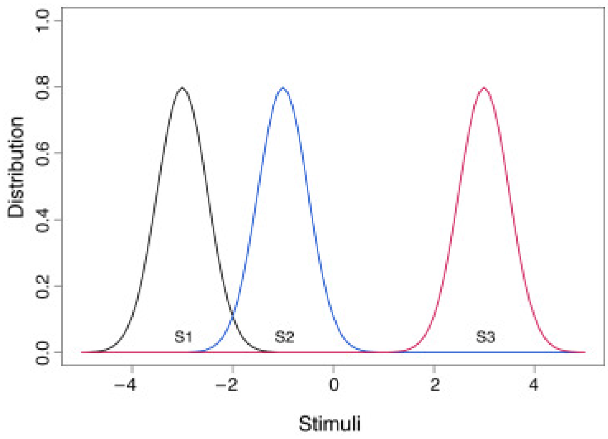
$$\mathbb{P}\left(\{\omega \in \Omega \mid r_S(c, \omega) = i\}\right).$$

3.2.2 Properties

The mathematical foundation of ranking models has a long history in different fields of science. Thurstone (1927) established a framework, which was later refined and extended by other researchers. The so-called Thurstone V models[51] represent a class of models in which items create a stochastic stimulus and are ranked accordingly. When correlation is set to zero, the stimulus is drawn independently from identical distributions, which only differ about their mean, see Figure 3.1 for an example.

[51] The Roman number V is a numeration of increasingly structured models.

Figure 3.1: Thurstone V model, where stimuli of items S1, S2 and S3 are drawn from normal distributions that differ about their mean. Subsequently, items are ranked according to their stimuli.



Bradley and Terry (1952) established a Thurstone V model with underlying extreme value distributions, which is the foundation of logit analysis known today. While the authors focus on pairwise comparisons, this approach was extended by Luce (1959) for rankings of n alternatives. In this work, he postulated what is today known as *Luce's Choice Axiom*,

$$p_S(x) = \frac{p_C(x)}{\sum_{y \in S} p_C(y)}. \tag{3.1}$$

Here, $p_S(x)$ denotes the probability of item x being ranked first when the set S is to be ranked,

$$p_S(x) = \mathbb{P}\left(\{\omega \in \Omega \mid r_S(x, \omega) = 1\}\right)$$

Intuitively speaking, LCA demands that the probability of being ranked first in a set equals the share of winning probabilities in the entire world. Note that it is crucial to define the relation on the probabilities to be ranked first. Reversing Luce's Choice Axiom in a way that "the probabilities of being ranked last in a set are caclulated using the probabilities of being last in the complete set of items" is a different assumption, and Luce himself proved that there is no distribution satisfying both his Choice Axiom and the "Reversed Choice Axiom" (see Luce (1959)).

Furthermore, he introduced the property of *Positivity* which is defined as

$$p_S(x) > 0 \qquad \forall S \subseteq C, \;\; \forall x \in S. \tag{3.2}$$

and states that each item has a positive probability to be chosen in any given choice set. Actually, one might be inclined to say that this property rules out the existence of any "irrelevant" option in any set. Specifically, as Luce (1959) pointed out himself, in any situation where LCA holds, $p_S(x) = 0$ for arbitrary S means that item x is never chosen in *any* set and could thus be eliminated.

Finally, Luce (1959) introduced the property of *Independence of Irrelevant Alternatives* (IIA) which, under the assumption of positivity, is formulated as

$$\frac{p_S(x)}{p_S(y)} = \frac{p_{S'}(x)}{p_{S'}(y)} \qquad \forall S, S' \subseteq C, \;\; \forall x, y \in S \cap S'. \tag{3.3}$$

Intuitively speaking, it states that the ratio of probabilities to be ranked first is constant for any two items in any two sets which contain both elements. It turns out that

Lemma 3.1. *The property of* IIA *is equivalent to* Luce's Choice Axiom *under the assumption of positivity.*

For a proof, see Breitmoser (2019). An alternative proof is provided in Appendix 3.5.1.

In the following, a novel independence property is defined. The property of *Independence of Alternatives* (IoA) assumes a certain structure on rankings, specifically on the probability of item x receiving a better rank than item y when ranking alternatives of the set S with $x, y \in S$. IoA states that this probability is independent of S.

Definition 3.1. *A model satisfies* Independence of Alternatives *if, for all $S, S' \in \Sigma$ and all $x, y \in S \cap S'$,*

$$\mathbb{P}\left(\{\omega \in \Omega \mid r_S(x,\omega) < r_S(y,\omega)\}\right) = \mathbb{P}\left(\{\omega \in \Omega \mid r_{S'}(x,\omega) < r_{S'}(y,\omega)\}\right).$$

Note that IoA and IIA are technically very different properties, as none implies the other. While IIA imposes a certain structure exclusively on probabilities to be ranked first, IoA relaxes this structure on rank one probabilities[52], but at the same time imposes additional structure on rank two to n. A simple example which satisfies IIA, but not IoA, as well as some further remarks can be found in section 3.2.3.

3.2.3 Relation between IIA and IoA

In order to see that IIA does not imply IoA, consider the following

Example 3.3. *Assume that the three alternatives $S = \{A, B, C\}$ are to be ranked. In pairwise comparisons, A is twice as likely to be first when compared to B or C, i.e.*

$$p_{\{A,B\}}(A) = \mathbb{P}\left(\{\omega \in \Omega \mid r_{\{A,B\}}(A,\omega) < r_{\{A,B\}}(B,\omega)\}\right) = \frac{2}{3},$$
$$p_{\{A,C\}}(A) = \mathbb{P}\left(\{\omega \in \Omega \mid r_{\{A,C\}}(A,\omega) < r_{\{A,C\}}(C,\omega)\}\right) = \frac{2}{3}.$$

In the comparison of B and C, both alternatives are equally likely to be first. Now, in the rating of all three alternatives, assume that

$$p_S(A) = \mathbb{P}\left(\{\omega \in \Omega \,| r_S(A,\omega) = 1, r_S(B,\omega) = 2, r_S(C,\omega) = 3\}\right) = 0.5,$$
$$p_S(B) = \mathbb{P}\left(\{\omega \in \Omega \,| r_S(B,\omega) = 1, r_S(A,\omega) = 2, r_S(C,\omega) = 3\}\right) = 0.25,$$
$$p_S(C) = \mathbb{P}\left(\{\omega \in \Omega \,| r_S(C,\omega) = 1, r_S(A,\omega) = 2, r_S(B,\omega) = 3\}\right) = 0.25.$$

One can easily verify that IIA holds, as A is still twice as likely to be first. On the other hand,

$$\mathbb{P}\left(\{\omega \in \Omega \mid r_S(A,\omega) < r_S(B,\omega)\}\right) = p_S(A) + p_S(C) = 0.75,$$

which is inconsistent with the pairwise comparison of A and B.

This example discloses what IIA lacks in order to imply IoA: an extension of its (strong) structure of probabilities for ranking first to the probabilities of other ranks. If additional structure is assumed, for example, probabilities to rank second being connected to the

[52] Consider, for example, the urn models in Chapter 1, which obviously satisfy IoA, while one of them does not satisfy IIA (Luce, 1959).

probabilities to rank first, IIA can imply IoA. One way of achieving this is by introducing *Order Consistency* (OC). To my knowledge, this axiom is new in the context of stochastic rankings.

Definition 3.2. *A model satisfies* Order Consistency *if* $\forall S \subseteq C$, $\forall i, j \in S$*:*

$$\mathbb{P}\left(\{\omega \in \Omega \mid r_S(x,\omega)=1,\ r_S(y,\omega)=2\}\right) = \mathbb{P}\left(\{\omega \in \Omega \mid r_{S\setminus\{x\}}(y,\omega)=1\}\right).$$

Intuitively speaking, OC means that there is no information to be gained about the ranking of other alternatives by observing the best option, i.e. correlation is set to zero. Now, it can be shown that

Lemma 3.2. *Together, the properties of* IIA *and* OC *imply* IoA*.*

For a proof, see Appendix 3.5.2. Note that OC is not a necessary condition for IoA to hold, i.e. one can find examples of four or more alternatives where IIA and IoA hold, but OC does not.
One might be inclined to generalize OC in a way that observing *any* rank does not change ranking probabilities of the other alternatives, i.e. $\forall S \subseteq C,\ \ \forall i, j \in S$:

$$\mathbb{P}\left(\{\omega \in \Omega \mid r_S(x,\omega)=i,\ r_S(y,\omega)=j\}\right) = \begin{cases} \mathbb{P}\left(\{\omega \in \Omega \mid r_{S\setminus\{x\}}(y,\omega)=j-1\}\right) & \text{if } i<j \\ \mathbb{P}\left(\{\omega \in \Omega \mid r_{S\setminus\{x\}}(y,\omega)=j\}\right) & \text{if } j<i \end{cases} \tag{3.4}$$

However, while this property might be intuitively intriguing, note that there is no model that can satisfy it (other than trivial models, where all alternatives have an equal chance in every set). The argument is as follows: property (3.4) implies both OC and IoA[53], hence LCA. At the same time, it implies "reversed OC", which together with IoA implies "reversed LCA". Thus, no non-trivial model can satisfy (3.4). One might say that this property is equivalent to generalizing Luce's axiom, extending it to probabilities for any rank, which is known not to be feasible.
The next section derives a result which holds solely under the assumption of IoA.

[53]See Appendix 3.5.3 for a formal proof.

3.3 Expected ranks from pairwise comparisons

The property of IoA in ranking models yields a relation between expected ranks in pairwise comparisons and expected ranks in sets of multiple alternatives. Let S be a set of k items to be ranked, i.e. $|S| = k$. Furthermore let $(\Omega, \mathcal{A}, \mathbb{P})$ be a general abstract probability space. Now, continue by defining the following random variables. Let $\hat{X}^S_{x,y}(\omega) : \Omega \to \{0, 1\}$ denote the random variable which (in state ω) yields a one if, in the overall ranking, x is ranked better than y, and zero otherwise.

$$\hat{X}^S_{x,y}(\omega) = \begin{cases} 1 & \text{for } r_S(x,\omega) < r_S(y,\omega) \\ 0 & \text{otherwise} \end{cases}$$

In addition, let $Y^S_x(\omega) : \Omega \to \{0, 1, ..., k-1\}$ be the random variable given by

$$Y^S_x(\omega) = \sum_{\substack{y \in S \\ y \neq x}} \hat{X}^S_{x,y}(\omega),$$

i.e. the random variable that states how many items have a higher rank than item x. Finally, recall that $r_S(x, \omega)$ denotes the rank of x.

Remark 3.1. *The rank $r_S(x, \omega)$ of item x when ranking set S equals (by definition) the difference between the total number of items and the number of items which are ranked higher than x, namely $Y^S_x(\omega)$:*

$$r_S(x,\omega) = k - Y^S_x(\omega) \tag{3.5}$$

Now, under the assumption of IoA, one can consider $\hat{X}^{\{x,y\}}_{x,y}(\omega)$, i.e. the random variable which yields a one if x is ranked better than y in the set $S = \{x, y\}$ (i.e., in the pairwise comparison of x and y) and claim that

$$\mathbb{P}\Big(\hat{X}^S_{x,y}(\omega) = 1\Big) \overset{(IoA)}{=} \mathbb{P}\Big(\hat{X}^{\{x,y\}}_{x,y}(\omega) = 1\Big) \tag{3.6}$$

From here, a link in expectation can be established, connecting the number of items that have a higher rank than x to the probabilities of pairwise comparison.

Lemma 3.3.

$$\mathbb{E}\left[Y_x^S(\omega)\right] = \sum_{\substack{y=1\\y\neq x}}^{k} \mathbb{P}\Big(\hat{X}_{x,y}^{\{x,y\}}(\omega)=1\Big) . \tag{3.7}$$

Proof. By definition, it holds that

$$\mathbb{E}\left[Y_x^S(\omega)\right] = \mathbb{E}\left[\sum_{\substack{y=1\\y\neq x}}^{k} \hat{X}_{x,y}^{S}(\omega)\right] .$$

Now, by (3.6) and noting that probabilities imply expected values, it holds that

$$\mathbb{E}\left[\sum_{\substack{y=1\\y\neq x}}^{k} \hat{X}_{x,y}^{S}(\omega)\right] = \mathbb{E}\left[\sum_{\substack{y=1\\y\neq x}}^{k} \hat{X}_{x,y}^{\{x,y\}}(\omega)\right]$$

and subsequently,

$$\mathbb{E}\left[\sum_{\substack{y=1\\y\neq x}}^{k} \hat{X}_{x,y}^{\{x,y\}}(\omega)\right] = \sum_{\substack{y=1\\y\neq x}}^{k} \mathbb{E}\left[\hat{X}_{x,y}^{\{x,y\}}(\omega)\right] = \sum_{\substack{y=1\\y\neq x}}^{k} \mathbb{P}\Big(\hat{X}_{x,y}^{\{x,y\}}(\omega)=1\Big) .$$

□

Combining (3.5) and (3.7) yields the following

Theorem 3.1. *Under IoA, the expected rank of item i in the ranking of k alternatives is determined by it's pairwise comparisons with all other alternatives, i.e.*

$$\mathbb{E}\left[r_S(x,\omega)\right] = k - \sum_{\substack{j=1\\j\neq i}}^{k} \mathbb{P}\Big(\hat{X}_{x,y}^{\{x,y\}}(\omega)=1\Big) .$$

The value of this relation lies in its connection between pairwise comparison and ranking of $k > 2$ items. The left-hand-side corresponds to

$$\mathrm{E}[r_S(x,\omega)] = \sum_{j=1}^{k} j \cdot \mathbb{P}\left(\{\omega \in \Omega \mid r_S(x,\omega) = j\}\right)$$

where $\mathbb{P}\left(\{\omega \in \Omega \mid r_S(x,\omega) = j\}\right)$ might be remarkably more complex to calculate.

The differences between the concepts and implications of IIA and IoA might best become apparent when considering chess ratings. Both the United States Chess Federation (USCF) and the Fédération Internationale des Échecs (FIDE) use rating systems to measure the strength of players and to predict outcomes. Notably, their formulas to calculate predictions differ. The USCF formulas correspond to a Thurstone V model with underlying extreme value distributions. On the other hand, FIDE assumes a Thurstone V model with underlying normal distributions. Yellott Jr (1977) shows that LCA (and therefore IIA) hold in the former, but not in the latter. Now, consider a chess tournament with multiple players. According to the results of Luce (1959) and McFadden (1973), one can infer the winning probabilities of the whole tournament from pairwise comparisons in the USCF system, but not in the FIDE system. On the other hand, as IoA holds, expected ranks can be derived from pairwise comparisons in both systems.

3.4 Conclusion

This paper motivates to consider the notion of IoA, which in some sense might seem competing with the notion of IIA. However, while their semantic appearance is similar, their formal definitions as well as their implications differ significantly. Debreu's critique about LCA and the recently discovered link between IIA and LCA demands for a closer re-evaluation. Intuitively speaking, one might wonder why random utility models such as Thurstone V as well as certain urn models would violate IIA. Potentially it could be fair to say that models which keep the stimulus of items constant (independent from other alternatives in the choice set), or models in which the reversed LCA holds, satisfy some independence property. The notion of IoA is in line with Luce's intuitive description of independence and could fix this issue, while at the same time allowing a cleaner interpretation of models that do or do not satisfy an independence property. In applications (such as chess tournaments), one might wonder whether some independence property holds, i.e. whether the level of performance is unaffected by other players who might or might not participate in the tournament. While there potentially can be arguements both for and against this property, it seems unintuitive to say that independence of irrelevant alternatives could hold in the USCF system, but (by definition) not in the FIDE system. As IoA holds in both, this issue does not exist when discussing independence of alternatives in chess tournaments.

In addition, this paper shows that, under the assumption of IoA, expected ranks can be determined solely from pairwise comparisons. This result holds specifically, but not exclusively, for Luce's model with underlying extreme value distributions. It constitutes a shortcut to calculate expected ranks in practice when the standard way of calculation could be significantly more cumbersome. The benefits of calculating expected ranks from pairwise comparisons might also simplify situations in which the comparison of multiple alternatives is significantly more difficult to comprehend.

3.5 Appendix

3.5.1 Proof of Lemma 1

Proof. First, note that positivity is necessary to guarantee that probability ratios are well-defined, i.e. none of the denominators is zero. Assuming that Luce's Choice Axiom (3.1) holds,

$$\frac{p_S(x)}{p_S(y)} \overset{(3.1)}{=} \frac{\dfrac{p_C(x)}{\sum_{z\in S} p_C(z)}}{\dfrac{p_C(y)}{\sum_{z\in S} p_C(z)}} = \frac{p_C(x)}{p_C(y)} = \frac{\dfrac{p_C(x)}{\sum_{z\in S'} p_C(z)}}{\dfrac{p_C(y)}{\sum_{z\in S'} p_C(z)}} \overset{(3.1)}{=} \frac{p_{S'}(x)}{p_{S'}(y)}.$$

On the other hand, assuming that IIA (3.3) holds and taking $S' = C$ yields

$$p_S(x) = p_C(x) \cdot \frac{p_S(y)}{p_C(y)}. \tag{3.8}$$

Summing (3.8) over all $x \in S$ yields

$$\sum_{x \in S} p_S(x) = 1 = \frac{p_S(y)}{p_C(y)} \sum_{x \in S} p_C(x)$$

and therefore,

$$p_S(y) = \frac{p_C(y)}{\sum_{x \in S} p_C(x)}.$$

□

3.5.2 Proof of Lemma 2

Proof. Assume IIA and OC hold and show that IoA holds by induction over $|S|$. Obviously, if $|S| = 2$, IoA holds. Now, assume it holds for $|S| = n$. Consider a set with $n+1$ alternatives, and rank one option to be in first place. This leads to three cases:

(i) x is ranked first

(ii) y is ranked first

(iii) some z different from x, y is ranked first

Consider cases (i) and (ii). Conditional on x or y being ranked first, it holds that

$$\begin{aligned}
\mathbb{P}\left(\{\omega \in \Omega \mid r_S(x,\omega) < r_S(y,\omega)\}\right) &= \frac{p_S(x)}{p_S(x) + p_S(y)} = \frac{1}{1 + \frac{p_S(y)}{p_S(x)}} \overset{\text{(IIA)}}{=} \frac{1}{1 + \frac{p_{\{x,y\}}(y)}{p_{\{x,y\}}(x)}} \\
&= \frac{p_{\{x,y\}}(x)}{p_{\{x,y\}}(x) + p_{\{x,y\}}(y)} = p_{\{x,y\}}(x) \\
&= \mathbb{P}\left(\{\omega \in \Omega \mid r_{\{x,y\}}(x,\omega) < r_{\{x,y\}}(y,\omega)\}\right).
\end{aligned}$$

Furthermore, in case (iii), we are left with n alternatives to rank from two to n. Given that OC holds, the probabilities to be ranked second in this set correspond to those of being ranked first in the set without z. This means, that both OC and IIA hold in this set of n alternatives, and therefore by induction assumption, $\mathbb{P}\left(\{\omega \in \Omega \mid r_S(x,\omega) < r_S(y,\omega)\}\right) = \mathbb{P}\left(\{\omega \in \Omega \mid r_{\{x,y\}}(x,\omega) < r_{\{x,y\}}(y,\omega)\}\right)$. Combining these arguments for the different cases finalizes the proof. □

3.5.3 Generalized order consistency proof

Proof. First, note that OC is a special case of property (3.4). Hence, it is left to show that the property implies IoA. Consider two alternatives, x and y and their order in pairwise comparison. Now note that, due to property (3.4), the addition of new elements to the set will not change the relative order of x and y, which implies that

$$\mathbb{P}\left(\{\omega \in \Omega \mid r_S(x,\omega) < r_S(y,\omega)\}\right) = \mathbb{P}\left(\{\omega \in \Omega \mid r_{\{x,y\}}(x,\omega) < r_{\{x,y\}}(y,\omega)\}\right)$$

for arbitrary S. Therefore, IoA holds. □

Chapter 4

Does mining fuel bubbles? An experimental study on cryptocurrency markets

4.1 Introduction

Speculative bubbles are a major destabilizing factor for the economy and often imply severe real consequences (see Brunnermeier and Schnabel (2016) for a comprehensive review). There are many episodes of bubbles and crashes in history such as the tulip bubbles, the south sea bubbles and the Mississippi bubbles. However, bubbles in cryptocurrency markets dwarf any major historical bubbles in terms of magnitude and duration (Cheah and Fry, 2015; Bianchetti et al., 2018). Cryptocurrencies were originally devised as a communication protocol that facilitates decentralized electronic payments (Böhme et al., 2015) but are increasingly recognized as an investment vehicle (Glaser et al., 2014). At its peak price in 2017, Bitcoin constituted a market capitalization of $238 billion.[54] At the same time, they differ from traditional revenue generating assets (stocks or bonds), as no income such as dividends or interest will be paid while holding them. Cryptocurrencies should also not be simply viewed as conventional commodities, because they are intangible and have no actual usage in producing other products. It is perhaps the best to

[54] https://www.statista.com/statistics/377382/bitcoin-market-capitalization/

recognize them as a completely novel asset class (Burniske and White, 2017; Hong, 2017; Kristoufek, 2015). In practice, the majority of traders consider holdings of Bitcoin to be a speculative investment rather than a substitute for cash (Yermack (2015), Baur et al. (2018)). As more investors hold cryptocurrencies in their portfolios, the risk of speculative bubbles in these markets may contagion to other financial markets and subsequently the real economy (Manaa et al. (2019), Guo et al. (2011)). Thus, policy makers and financial institutions need to better understand the functioning of the cryptocurrency market in order to design regulation and help contain potential systemic risks.

Among all major cryptocurrencies, the original and most dominating one is Bitcoin. The concept of Bitcoin was introduced by Nakamoto et al. (2008) in a whitepaper as a decentralized peer-to-peer electronic cash system which uses a public protocol called blockchain to track ownership and transactions of tokens. The validity of entries into the public protocol as well as the generation of tokens (i.e. mining) is achieved using the proof-of-work algorithm. This means that, in order to add information to the protocol, a mathematical problem has to be solved.[55] The difficulty of these problems may vary over time, as the system is designed to self-adjust in order to limit the total amount of Bitcoin tokens that can ever be created. Importantly for our purposes, the difficulty (which translates into actual costs) has been prohibitively high for individual miners in recent years, fostering the rise of professional miners. Professional miners have dedicated equipment to efficiently mine major cryptocurrencies, while investors most likely can only purchase Bitcoins on the crypto market to include them in portfolios. A comprehensive description of proof-of-work and Bitcoin mining can be found in Auer (2019).

We identify several unique features of Bitcoin that are not shared by other asset classes. First, most cryptocurrencies use the proof-of-work method[56] and thus their tokens are costly to acquire, requiring miners to contribute non-trivial and ever-increasing computational power.

Second, a supply-smoothing (sticky supply) feature is typically built-in cryptocurrency

[55] The proof-of-work algorithm is implemented using *hash functions*, which means that obtaining a solution to the mathematical problem takes effort, while the verification of a solution is immediate.

[56] According to Gervais et al. (2016), Proof of Work (PoW) powered blockchains accounted for more than 90% of the total market capitalization of existing digital cryptocurrencies in 2016.

protocols. The mining difficulty is adjusted to ensure that the number of coins being generated in a given time window stays roughly constant, regardless of the number of miners at work (for details see: Nakamoto et al., 2008). When a demand shock occurs, the supply is sluggish, forcing upward pressure on prices. Third, market participants may envision that cryptocurrencies will eventually be widely accepted as a medium of transaction and thus become more valuable since the total amount of tokens is limited.[57] While the property of costly mining arguably shares similarities to resource extraction models, it is worth noting that cryptocurrencies do not depreciate after usage, and the speed of extraction is not dependent on the miners themselves.

We design a laboratory experiment to identify the characteristics that contribute to bubbles and crashes in cryptocurrency markets. Attaching objective probabilities to the possible outcomes of the current Bitcoin enthusiasm seems impossible (Shiller, 2017). This raises the question of how to model the value of Bitcoin in the lab. As our focus is on the proof-of-work method and whether its presence and accessibility has any impact on pricing behaviour, we favour a simple and clear design of our asset value. While Bitcoin enthusiasts are convinced that bitcoin has intrinsic value, their beliefs regarding the specific value can differ. We implement our lab version of a cryptocurrency as an asset with uncertain value. We examine asset markets which only differ in the way assets are provided, either in a standard endowment market where assets are "a gift", or in a proof-of-work (mining) type market. We study these two paradigms both with and without 'barriers' to the market. In the gift paradigm, only half of the traders receive an initial stack of assets as a gift but everyone can trade. In the mining paradigm, only half of the traders can mine and trade, while the rest are only allowed to trade.

We are the first to design a controlled laboratory environment to study cryptocurrency markets. Krafft et al. (2018) conduct an online experiment, inducing buy activity in crypto markets by using bots that trade only pennies, and pointing out the potential impact of peer influence effects. Gurdgiev and O'Loughlin (2020) apply sentiment analysis to

[57] Clearly, the properties of constant supply and an upper-limit to total supply appear to be contradicting each other. In the case of Bitcoin, the supply is kept constant for an approximate time span of two years. Subsequently, the supply for the next two years is halved.

publicly accessible data and find indications that herding and anchoring biases influence cryptocurrency prices. While these papers reveal potential behavioral pattern in crypto markets, they do not focus on explanations for the overpricing observed in cryptocurrency markets. This is the main objective of our study.

The research of bubbles and crashes in financial markets has a long history in experimental economics. Smith et al. (1988) observe systematic overpricing across various treatments. These results lead to a large number of related studies, which Palan (2013) and Powell and Shestakova (2016) comprehensively review. Our design shares similarities with Smith et al. (1988), but specifically differs in dividend payments. In our market, assets pay no repeating dividends in every period. Instead, the value of the asset at the end of the trading periods is randomly drawn. This is equivalent to a one-time dividend payment at the end of the last period, which we call *redemption value*. This implies that throughout the trading periods the assets have a constant expected value (the fundamental value is flat). There are some existing studies that investigate determinants of price bubbles for assets with a flat fundamental value. Noussair et al. (2001) and Holt et al. (2017) implement a design with dividends in every period, but constant fundamental value. They find that bubbles are smaller but persist compared to decreasing fundamental value markets. Cueva and Rustichini (2015) conduct experiments with flat fundamental value in order to investigate gender effects in financial markets. Overall, they find little overpricing. Noussair and Tucker (2016) highlight how cash holdings influence pricing in flat fundamental value markets; they find that larger initial cash endowments facilitate overpricing. Kirchler et al. (2012) find that in markets with constant fundamental value, prices track fundamental value well. Overall, the existing literature seems to suggest that no large bubbles or crashes should be expected in our gift treatments. However, to the best of our knowledge we are the first to implement a design where traders acquire their assets costly (our mining treatments). Whether overpricing for a flat fundamental value asset is observed or not is thus unclear in such a design given the existing literature. We envision our contribution helps link asset market better with real-world applications, as put forward by Oechssler (2010).

The main contribution of this paper is to offer a systematic evaluation of the intertwining factors that may be attributed to bubbles in cryptocurrencies. We look into the effect of mining and the accessibility of mining technology on pricing, while isolating other potentially confounding factors. Our results show that, in the absence of mining, prices track fundamental value well. Yet, we find that the proof-of work algorithm enables overpricing, while combined with entry barries exacerbates it.

Section 4.2 describes our experimental design and procedures in detail. In section 4.3, we report our experimental results and section 4.4 concludes.

4.2 Design and Procedure

4.2.1 Experimental Design - Asset Market

We design an asset market experiment, where participants trade assets using experimental currency units (ECU). Assets have flat fundamental value, they only pay out an uncertain dividend at the end of trading, we call this the *redemption value*.

Our experiment features a 2x2 factorial design, summarized in Table 4.1. We vary how traders receive assets to trade; participants are either endowed with assets at the outset of the market, as a gift, or can decide to mine for assets at a cost. The costly mining incorporates both the sticky and the limited supply features of proof-of-work algorithm based cryptocurrencies. Note that costly mining implies that the cash-to-asset ratio (c/a ratio) in our mining treatments varies over time, we elaborate further on this below. Additionally, we vary if all traders have the same capacities in the market. In our gift treatment traders differ on whether they initially receive asset gifts or not, while in our mining treatment, traders differ in their access to mining. This yields four different treatments, namely assets as a gift for all (Gift-All), mining for all (Mining-All), assets as a gift for half (Gift-Half) and mining for half (Mining-Half).

Table 4.1: Summary of treatments.

	All	*Half*
Gift	Gift-all	Gift-half
Mining	Mining-all	Mining-half

Gift-All is our baseline condition in which all traders are endowed with an equal amount of ECUs and assets. At the end of the experiment, the asset holdings are bought back by the experimenter at a random value that is drawn from a skewed distribution and all ECU earnings are converted to Euros. The buyback value of the asset is a random redemption value of 0, 15, 30, or 67 with an equally likely chance, yielding an expected value of 28 ECUs. The flat but uncertain fundamental value captures the plausibly divergent views on how cryptocurrencies are valued by different investors.

In Gift-Half, we randomly assign half of the traders to be endowed with both assets and ECUs, while the other half does not receive any assets from the outset, but some additional ECUs.

Condition Mining-All is the same as Gift-All, except that assets do not come as a gift, but have to be mined. Participants are endowed with initial ECUs, but not initial assets. They have the option to spend ECUs in each period to obtain (mine) assets. We introduce this costly mining process, which resembles the proof-of-work algorithm, to identify whether mining initiates and/or exacerbates bubbles. The sticky and limited supply features identified earlier are incorporated here. Even when mining is potentially profitable (for example if assets are traded at prices higher than mining costs), there is a cap on how many assets traders can mine per period. Mining operates concurrently with the asset market. By contrasting Gift-All and Mining-All, we separate the effect of costly mining on asset pricing.

Condition Mining-Half is designed to capture the way cryptocurrency mining operates in the real world. For most cryptocurrencies, mining requires dedicated machines. This implies that many investors have no option to efficiently mine coins and are thus constrained to only trading them if they wish to obtain some. We study how asset pricing is affected when only half of the traders have the possibility to mine for assets, while the other half is restricted to acquire assets through trading in the market. With this treatment, we can identify how concentration of access to the mining technology influences the asset pricing.

The parameters of our baseline Gift-All are in accordance to the ones implemented by Weitzel et al. (2018). Traders are initially endowed with 5700 ECUs and 20 units of the asset. In Mining-All, traders begin with 5900 ECUs and zero assets. They can decide to spend up to 40 ECUs per period on mining. In Gift-Half and Mining-Half, traders are endowed according to their randomly assigned capabilities (role). Specifically, we implement two roles and randomly assign half of the traders to role A, while the rest to role B. In treatment Gift-Half, traders of role A are endowed with 40 assets and 5140 ECUs at the outset, while role B players are endowed with 6260 ECUs but no assets. Note that, given the expected redemption value of 28, the initial portfolios of traders in Gift-Half are equivalent to those of traders in Gift-All in terms of expected value. In Mining-Half, role A has a starting endowment of zero assets, 5540 ECUs and is allowed access to the mining technology. Role A traders can potentially spend up to 80 ECUs on mining in each period.[58] Instead, in the Mining-half treatment, role B traders are endowed with 6260 ECUs, no assets and have no access to the mining technology. For an overview of the parameters for each treatment arm, see Table 4.2.

Table 4.2: Overview of parameters across treatments

		All	Half	
			Role A	Role B
Gift	ECUs	5700	5140	6260
	Assets	20	40	0
Mining	ECUs	5900	5540	6260
	Assets	0	0	0
	Cap/Period	40	80	0

Special attention is given to the calibration of the experimental parameters to make conditions comparable across sessions. While the cash to asset ratio in Gift-All and Gift-Half is constant throughout the trading periods, it varies over time for the mining treatments (it is strictly decreasing whenever mining takes place). We aim to calibrate the paramaters in a way that the cash to asset ratios of gift and mining treatments are similar. Assume every trader in Mining-All mines the maximum amount (40 ECUs) in the first five periods, which is potentially a reasonable thing to do given the low cost of mining as compared to the fundamental value of the asset. If no other trading is performed in the meantime, their holdings would be 5700 ECUs and approximately 20 assets (exactly 19.3) at the end of period 5, which is approximately the initial endowment of traders in Gift-All. From period 6 onwards, the mining cost would exceed the expected value of

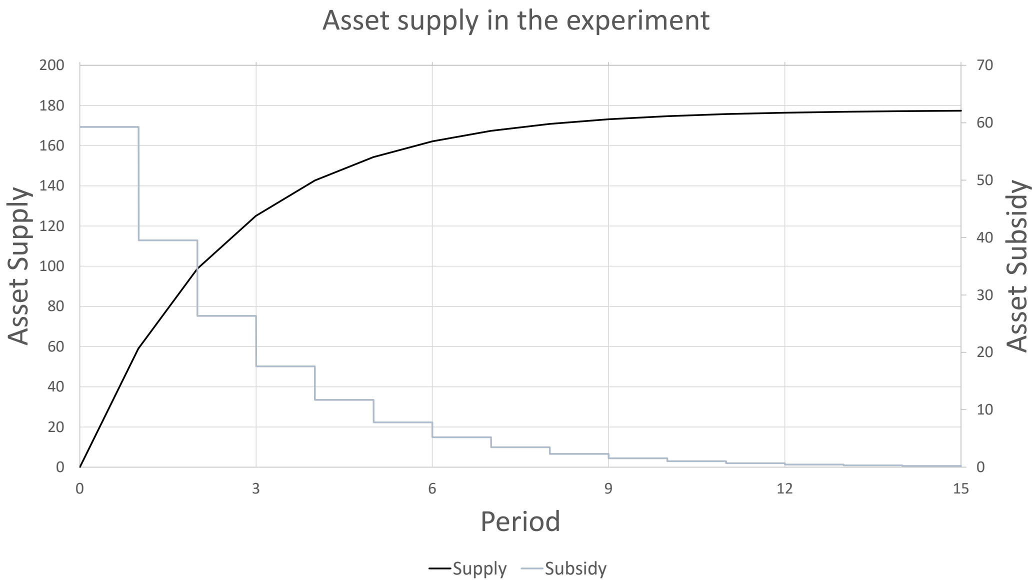
[58]We double their cap to allow for the same total mining volume as in Mining-All.

the asset, thus risk-neutral agents would refrain from further mining.[59] Analogously, if all role A traders in Mining-Half were to mine using the maximum amount (80 ECUs) in each of the first five periods, they would reach (approximately) the initial endowment of role A traders in Gift-Half.

The function to calculate the mining costs is the same for both Mining-All and Mining-Half treatments:

$$C\left(\sum_{i\in I,\ t<\hat{t}} x_{i,t}\right) = C\left(\chi_{\hat{t}}\right) = 5.4 \cdot 1.5^{\frac{\chi_{\hat{t}}}{40n}} \tag{4.1}$$

Figure 4.1: Asset supply in our mining treatments over time (assuming maximum mining).



where n denotes the number of traders in the market, $\hat{t}$ denotes the current period, and $x_{i,t}$ the mining expenditure of subject i in period t. Intuitively speaking, mining costs start at a level of 5.4 ECUs and increase by 50% in every period, given mining takes place at maximum effort.[60] Figure 4.1 depicts how asset supply in our mining treatments

[59] In this example, mining costs would increase from 27.3 to 41 ECUs per asset.

[60] We choose to update costs as a function of total expenditure, as we consider it easier to comprehend when explaining it to participants. Traders have a maximum that they are allowed to spend on mining in each period. Thus, this form allows immediate interpretation of asset costs over time (i.e., an increase by 50% in every period where people mine the maximum). Interpreting costs as a function of assets would make it necessary to calculate how many assets are actually affordable to see what the cost of the next period is. For completeness, Figure 4.8 in the supplementary material depicts costs as a function of assets in our experiment.

evolves over time. Comparison with the exponentially decreasing supply of Bitcoin over time (see Figure 4.2) indicates that our modelling of bitcoin supply is appropriate. We calibrate the cost function to entail mining costs at approximately the level of the asset's fundamental value after a third of the trading periods are completed, assuming all traders mine. In fact, traders were mining slightly slower than the we anticipated. Figure 4.3 depicts the cash to asset ratio evolution over trading periods in the four treatments.[61]

Figure 4.2: Bitcoin supply over time (www.coindesk.com).

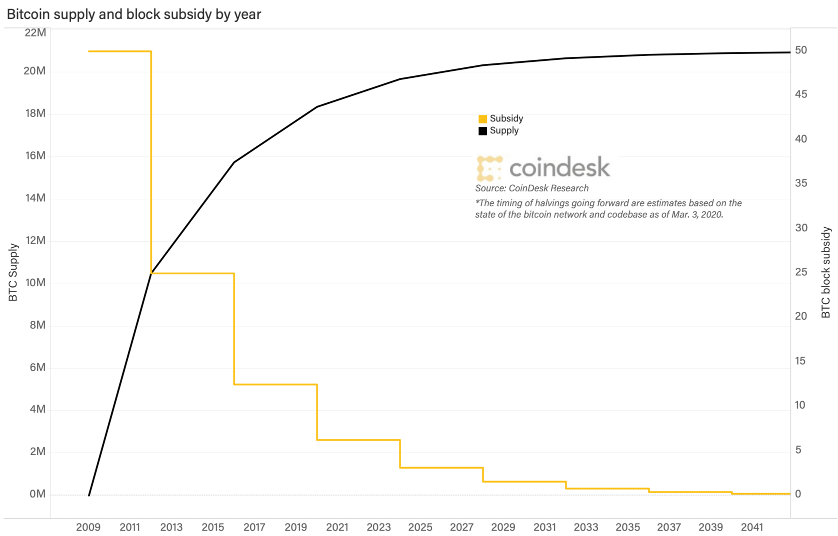
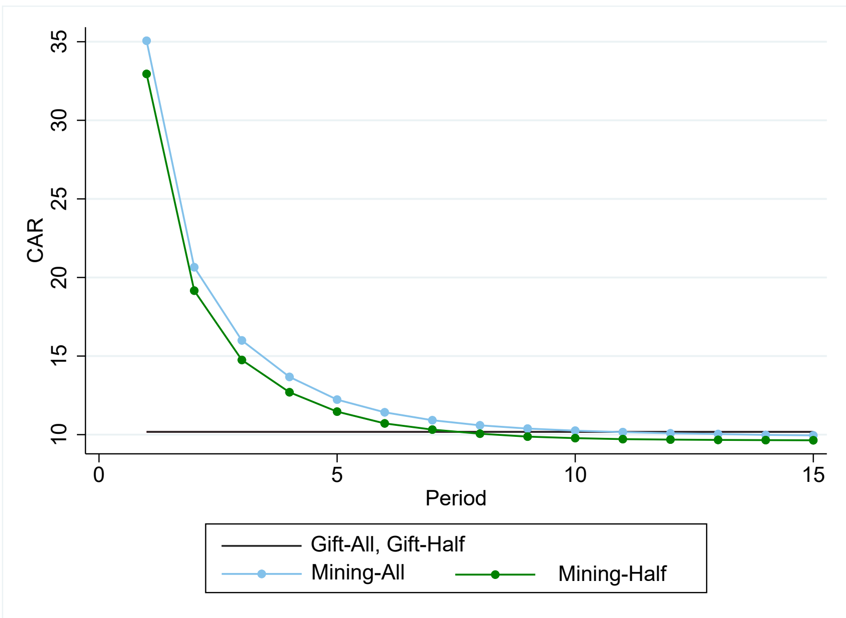


Figure 4.3: Average cash to asset ratio in all treatments



[61]The figure reports the average cash to asset ratio over the 9 markets implemented for each of the four treatments.

4.2.2 Experimental Design - Additional Measures

Before the asset market was administered, across all four treatments we elicit a number of individual traits and characteristics. For cognitive ability measurement, the participants complete a Raven Advanced Progressive Matrices (APM) test of 12 matrices. They have a maximum of 10 minutes for this task. The subjects are initially shown an example of a matrix with the correct answer provided below for 30 seconds. For each item a 3×3 matrix of images is displayed on the subjects' screen; the image in the bottom right corner is missing. The subjects are then asked to complete the pattern choosing one out of 8 possible choices presented on the screen. The 12 matrices are presented in order of progressive difficulty as they are sequenced in Set II of the APM. Participants are allowed to switch back and forth through the 12 matrices during the 10 minutes and change their answers. The subjects are rewarded with 1 Euro per correct answer from a random choice of two out of the total of 12 matrices. The Raven test is a non-verbal test commonly used to measure reasoning ability and general intelligence. Matrices from Set II of the APM are appropriate for adults and adolescents of higher average intelligence.

Following Bruguier et al. (2010), we also obtain a measure of Theory of Mind (ToM) using the Heider-Simmel task, a display of geometric shapes whose movements imitate social interaction (Heider and Simmel, 1944). We paused the movie every five seconds and asked subjects to predict whether two of the shapes would get closer or not within the next five seconds, rewarding them for correct predictions.

Finally, we elicit risk preferences using an incentivised Eckel and Grossman (2008) task.

Once the asset market was completed, we administer a questionnaire for general demographics, comprehension of the expected value of the asset traded and previous experience with cryptocurrencies.

These additional measures we obtain can help us identify possible characteristics of different traders and how traits might assist or hinder success in the market. We summarize participant characteristics by treatment and role in Table 4.3. Overall, our treatments are balanced, in particular with respect to gender, which is important to note given the findings of Eckel and Füllbrunn (2015).

Table 4.3: Characteristics of participants across treatments

	Gift-All	Gift-Half		Mining-All	Mining-Half	
		A	B		A	B
Markets	9	9		9	9	
Subjects	72	36	36	70	36	36
Avg. Age	23.54	22.28	24.94	24.21	21.75	22.72
Avg. Gender	0.58	0.39	0.47	0.47	0.58	0.47
Avg. Crypto Exp.	2.72	2.84	2.91	2.73	2.67	2.92
Avg. Raven	8.22	7.78	7.83	8.11	7.61	7.78
Avg. ToM	3.36	3.50	3.56	3.61	3.58	3.56
Avg. Risk Choice	3.54	3.54	3.31	3.51	3.42	3.81

4.2.3 Experimental Procedure

Eight traders participated in each market implemented across the four treatments, except for two sessions due to no-shows. The trading platform was implemented using Ztree (Fischbacher, 2007) and the technical toolbox GIMS (Palan, 2015). Trading was processed according to a continous double auction protocol, where traders can freely submit bids and asks. Every market consisted of 15 periods of 120 seconds each, and the holdings of assets as well as ECUs carried over from period to period. Shorting assets or borrowing ECUs was not possible. To determine the redemption value of our assets, we implemented a transparent randomization process which guaranteed that each of the four buyback values were assigned to exactly two participants.[62] This was done by having each trader physically draw from a deck of cards. The deck of cards had 4 pairs of cards. Each pair corresponded to one of the four buyback values. The cards were drawn without replacement by each of the 8 traders.

Our experimental sessions took place in economics lab facilities in Heidelberg University and Frankfurt University. Participants were mostly undergraduate students from a variety of majors. We implemented nine sessions for each of the four conditions. A total number of 286 subjects participated, using ORSEE (Greiner, 2015) in Frankfurt and SONA (www.sona-systems.com) in Heidelberg.

Prior to the asset market, all subjects were handed identical versions of experimental instructions. We conducted a quiz which every participant had to pass after reading

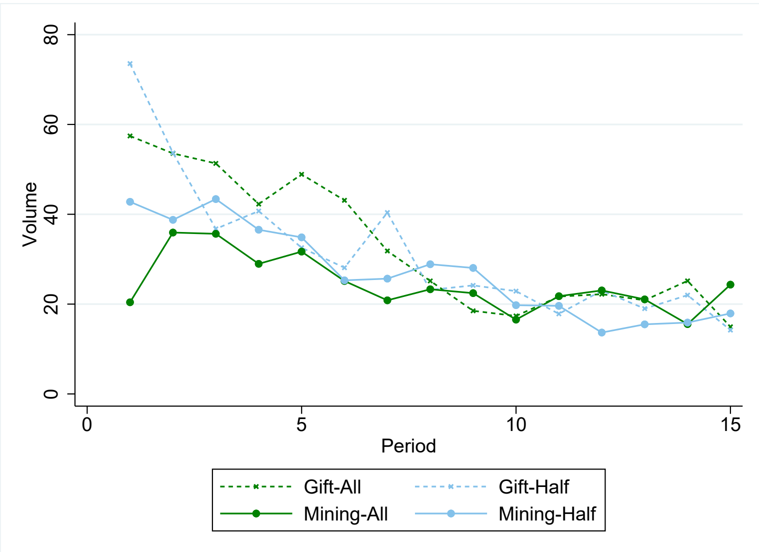
[62] In the sessions with only seven participants, one of the buyback values was assigned to only one participant, and which of the values would be assigned only once was part of the random procedure.

the instructions, asking about features and parameters of the experiment. Before the beginning of the 15 market periods, each session was allowed to practice for three periods of 120 seconds each. The asset and ECU holdings were reset after these practice rounds (practice rounds did not count towards final earnings). In Gift-Half and Mining-Half, the roles of traders were randomly determined before the practice rounds and their role remained the same after the practice. The average payment was approximately 18 Euro for 90 minutes.

4.3 Results

The focus of our analysis is on the market level. That is, we describe the characteristics of a market rather than the characteristics of individual traders. Note that each of our four treatments consists of nine market observations.

Figure 4.4: Average trading volume per period in all treatments



We present the average trading volume of each treatment across periods in Figure 4.4. Given our parameter calibration and similar cash-to-asset ratios from the 6th period onwards, we do not expect trading volumes to differ across our treatments. The gift treatments appear to initial trade in higher volumes. A possible reason for this could be the fact that in the first few periods, there are substantially fewer assets available to trade in the mining treatment markets. Figure 4.3 shows the average Cash-to-Asset-

Ratio (CAR) of our treatments across periods. By comparing Figure 4.4 and Figure 4.3, it seems that trading volumes of all four treatments are not significantly different once the cash to asset ratio is similar (from 6th period onwards). This is confirmed using a non-parametric test of comparing average trading volumes of periods 6-15 across the four treatments (Wilcoxon rank-sum test of Mining vs Gift, $p = 0.39$; Wilcoxon rank-sum test of all vs half, $p = 0.80$). Thus, we summarise the results until here by:

Result 4.1: *Once cash to asset ratios are similar, trading volumes are no different across all four treatments.*

Figure 4.5: Weighted average price per period of the median session in all treatments

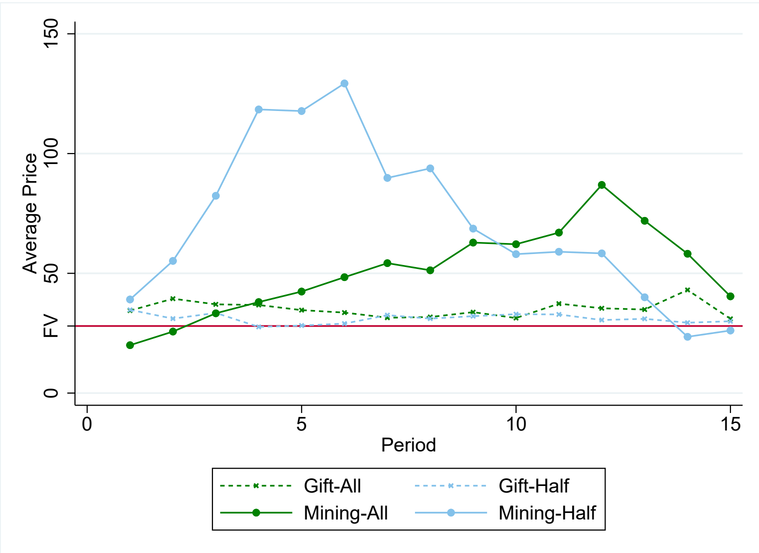


Figure 4.5 depicts the average trading price of the asset across the four different treatments. We report the median prices in each treatment based on volume weighted average prices from each market. Our median analysis is robust to outliers, in Figure 4.6 we replicate Figure 4.5 by systematically removing one of the 9 markets sequentially. The general tendencies and conclusions we make according to Figure 4.5 remain. We report weighted averages because traders are allowed to trade fractions of assets.[63] It is clear to see that in both Gift treatments trading prices track the asset fundamental value quite well through all trading periods. In other words, under both Gift treatments we observe no overpricing. In Mining-All, prices initially start below or very close to fundamental value but above the mining cost. Pricing then follows an upward trend parallel to the mining cost

[63]We provide graphs reporting averages instead of medians in the supplementary material.

with a clearly identifiable mark-up (see Figure 4.7 where we depict trading prices only for the two mining treatments together with their respective mining cost trends). Overall, prices rise slowly for about 12 periods, before they crash toward the fundamental value in the last three periods. Similarly, in treatment Mining-Half, prices go above and beyond fundamental value.

Figure 4.6: Robustness check: weighted average price per period of the median session of all but one session in all treatments, which yields eight graphs per treatment. We shade the area between the highest and lowest period prices per treatment, i.e. all eight graphs of a treatment lie within the shaded area of that treatment.

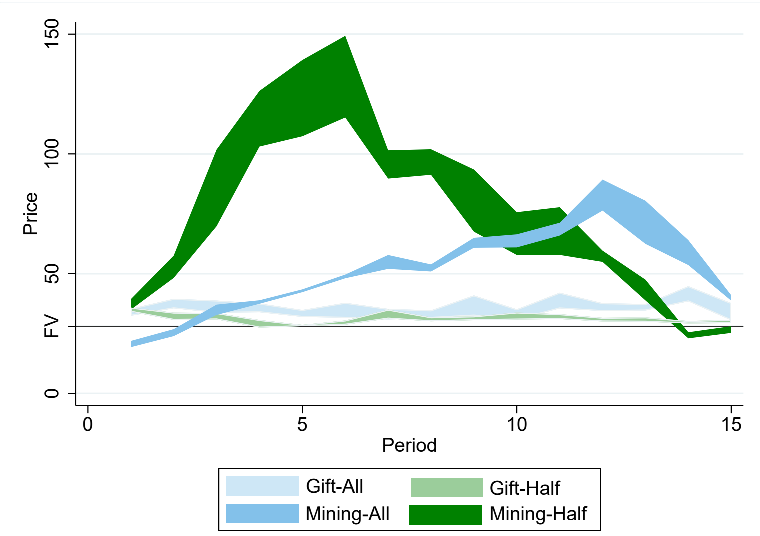
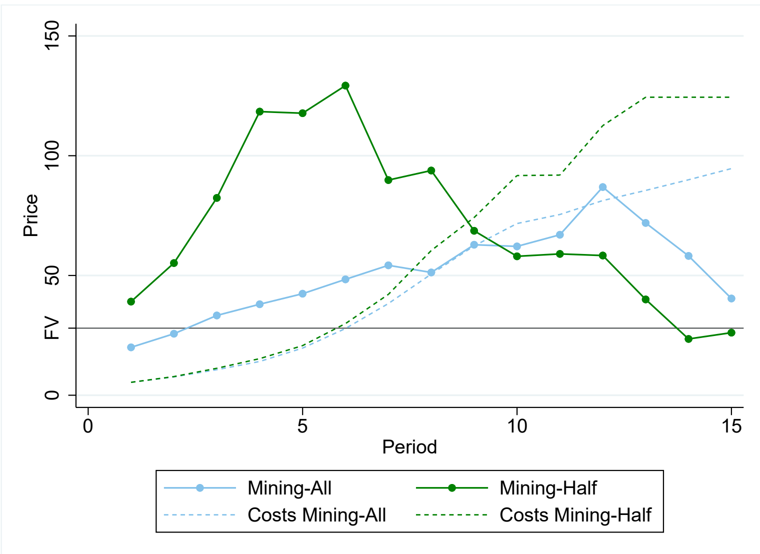


Figure 4.7: Weighted average price and mining cost per period of the median session in mining treatments



We formalize our results using a number of bubble measures also analysed by Weitzel et al. (2018), which are summarized in Table 4.4. These indicators include RD, the relative deviation of prices to fundamentals (normalized at the fundamental value of 28) and RAD, the relative absolute deviation of prices to fundamentals (normalized at the fundamental value of 28), both of which were introduced by Stöckl et al. (2010). RAD measures how closely prices track fundamental value, while RD indicates whether prices on average are above or below fundamental value. Furthermore we include RDMAX, measuring the overpricing of the peak period, AMPLITUDE, which captures the relative difference of the pre-peak minimum price and the peak price in terms of magnitudes of the fundamental value, and CRASH, which compares the peak price to the minimum price post-peak. RDMAX, AMPLITUDE and CRASH were established by Razen et al. (2017). Following Weitzel et al. (2018), we also compute the indicators TURNOVER, LIQUIDITY, SR (submission rate), SPREAD and VOLA (volatility). TURNOVER measures the volume of trade. LIQUIDITY describes the volume quantities of open orders at the end of each period, while SR is defined as the number of limit orders posted divided by the sum of limit and market orders posted in a period. SPREAD measures the gap between buy and sell orders, and VOLA measures log-returns of all market prices within a period.[64]

Table 4.4: Summary statistics of bubble measures by treatment

	Gift-All		Gift-Half		Mining-All		Mining-Half	
	mean	(std.dev.)	mean	(std.dev.)	mean	(std.dev.)	mean	(std.dev.)
RAD	0.5	(0.5)	0.5	(0.8)	1.9	(1.9)	2.3	(1.5)
RD	0.4	(0.5)	0.5	(0.9)	1.9	(1.9)	2.2	(1.5)
RDMAX	0.9	(0.6)	0.9	(1.4)	7.7	(10.7)	6.1	(5.2)
AMP	0.7	(0.3)	0.6	(0.6)	7.9	(10.7)	5.6	(5.0)
CRASH	-0.6	(0.6)	-0.7	(0.9)	-7.5	(11.3)	-6.2	(5.4)
TURN	0.2	(0.1)	0.2	(0.1)	0.2	(0.1)	0.2	(0.1)
LQ	0.8	(0.7)	5.5	(13.9)	0.7	(0.6)	5.2	(12.8)
SR	20.4	(4.4)	21.1	(4.9)	16.3	(3.5)	21.9	(3.5)
SPREAD	0.3	(0.2)	0.2	(0.3)	1.4	(2.3)	1.5	(1.2)
VOLA	0.3	(0.3)	0.2	(0.1)	0.3	(0.2)	0.5	(0.3)

The conclusions drawn regarding our four treatments from inspecting Figure 4.5 remain. Gift-All and Gift-Half are characterised by moderately sized bubble measures. Additionally, we find no significant difference when comparing the bubble measures of the two Gift treatments. When comparing the bubble measures of the mining treatments to their respective gift treatment (Gift-All vs. Mining-All & Gift-Half vs. Mining-Half), in Table 4.5, we find significant differences. Specifically, price deviations from fundamental

[64]For a more detailed description of all bubble measures see the supplementary material.

value are significantly more pronounced in the mining treatments as compared to the gift treatments. We summarise the comparison of mining and gift treatments by:

Result 4.2: *Asset prices in our Gift-treatments track fundamental value well. Overpricing in our Mining-treatments is significantly larger than in the respective Gift-treatments.*

Table 4.5: p-values of exact Mann-Whitney-U tests comparing bubble measures of different treatments (pairwise)

	Gift-All vs. Gift-Half	Gift-All vs. Mining-All	Gift-Half vs. Mining-Half	Mining-All vs. Mining-Half
RAD	0.546	**0.004*** **	**0.003*** **	0.666
RD	0.387	**0.006*** **	**0.004*** **	0.605
RDMAX	0.387	**0.001*** **	**0.002*** **	0.931
AMPLITUDE	1.079	**0.002*** **	**0.036** **	1.000
CRASH	0.673	**0.005*** **	**0.001*** **	0.606
TURN	0.863	0.931	0.387	0.546
LQ	0.340	1.000	1.000	0.297
SR	1.000	**0.050* **	0.666	**0.006*** **
SPREAD	0.340	**0.006*** **	**0.000*** **	0.136
VOLA	0.222	0.161	0.014	0.436

Table 4.6: Exact Mann-Whitney-U tests of bubble measure RAD, comparing first half and second half of mining treatments

		Mining-All	Mining-Half	Mining-All vs. Mining-Half (M-W-U test)
RAD	First half	0.60	2.20	p=0.011
	Second half	1.33	0.81	p=0.436
RD	First half	0.47	2.20	p=0.008
	Second half	1.33	0.81	p=0.340

As argued above, we are also interested in identifying what effect concentration of access to the mining technology might have on asset pricing. Thus, we now focus on contrasting our two mining treatments. We find no significant difference when comparing the bubble measures of Mining-All to Mining-Half (Table 4.5). However, inspection of Figure 4.7 suggests that a difference in the timing of the bubbles in Mining-All and Mining-Half exists. In the latter, prices seem to decouple from the mining cost already within the first few periods and peak even higher. Table 4.6 compares our mining treatments Mining-All and Mining-Half when we split the trading periods in two halves. We define periods $1-7$ as the first half and periods $9-15$ as the second half. This analysis confirms our observations made from Figure 4.7. The bubble measures RAD and RD of our mining treatments show a statistically significant difference in the first half of trading periods.[65] The peak price comes much earlier in treatment Mining-Half compared to Mining-All, and the bubble persists for a number of periods before prices crash to fundamental value. We summarise our result of comparing Mining-All and Mining-Half by:

Result 4.3: *The degree of overpricing of Mining-All and Mining-Half does not differ overall, but shows significant difference in the first half of the trading periods. In particular, Mining-Half markets tend to result in overpricing earlier than Mining-All markets.*

Table 4.7: Profit regressions with robust standard errors (reported in parentheses) clustered at the session level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Standardized earnings (Euro)	All (1)	All (2)	Mining-All, Mining-Half (3)	Gift-Half, Mining-Half (4)
Raven score	0.277** (0.088)	0.247** (0.099)	0.311** (0.105)	0.281*** (0.077)
Theory of Mind	0.383* (0.183)	0.344 (0.206)	0.451** (0.159)	0.512** (0.170)
Raven score X Theory of Mind	-0.458* (0.022)	-0.040 (0.025)	-0.063** (0.022)	-0.055** (0.018)
Controls				
Risk choice		0.064 (0.039)	0.022 (0.067)	0.058 (0.049)
Age		0.001 (0.098)	-0.012 (0.012)	0.005 (0.007)
Gender		-0.240* (0.111)	-0.472** (0.204)	-0.168 (0.187)
Crypto Exp.		0.040 (0.087)	-0.053 (0.084)	0.018 (0.060)
Role in Market				-0.371** (0.158)
Constant	-2.270** (0.687)	-2.261*** (0.646)	-1.759** (0.728)	-2.583** (0.868)
R^2	0.074	0.113	0.146	0.123
Observations	278	278	142	144

For the last part of our analysis, we study the characteristics of individual traders and how these might have an effect on earnings. Specifically, in Table 4.7, we report regression analysis of participant earnings. Inspired by Hefti et al. (2016) and Corgnet et al. (2018), we include an interaction term between cognitive abilities and Theory of Mind. The dependant variable in all specifications is the market earnings, standardized with respect

[65]Note that the bubble measures RDMAX, AMPLITUDE and CRASH are calculated with respect to the peak period and therefore can not be calculated when the sessions are split in half.

to their respective treatment. The first column reports the regression results without controls, while the second column includes control variables. In column 3, we report the regression results estimated only for our mining treatments. Finally, in column 4, we also control for the role of traders in the treatments Gift-Half and Mining-Half. Overall, we find that both cognitive ability and Theory of Mind are associated with higher earnings. These attribute appear to act as substitutes for each other as seen by the negative interaction term. Female traders seem to earn less, while in markets with different roles, traders that have no access to the asset other than from trading earn less.

4.4 Concluding Remarks

The price trajectories of cryptocurrencies, in partiular, Bitcoin, have raised questions among economists and policy makers. We design a laboratory experiment in order to look into the effect of mining and accessibility to the mining technology on asset pricing. The results of our experiment show a remarkable degree of overpricing, traders trade assets for significantly higher prices than the fundamental value. While risk preferences might explain slight overpricing, the maximum possible redemption value of any asset once trading is concluded is 67 experimental currency units. Nevertheless, we find that most mining markets far exceed this price. These findings indicate that the proof-of-work algorithm contributes to overpricing and enables significant volatility of pricing over time. The fact that we do not find overpricing in our Gift-treatments is in line with existing literature. First, a flat fundamental value (instead of a decreasing one) is simple and thus trading prices are more likely to adhere to fundamental value (Smith et al., 2000; Kirchler et al., 2012). Second, despite our relatively high cash to asset ratio, we do not pay frequent dividends as in Noussair et al. (2001). This observation supports their conjecture that in flat fundamental value settings, a high cash to asset ratio is not sufficient to ignite bubbles. On the other hand, Noussair and Tucker (2016) find large bubbles in flat fundamental value settings when frequent dividends are paid.

Restricted access to the mining technology ignites the overpricing further and pushes the prices to decouple from mining costs earlier. These results suggest that mining costs serve as a support of prices in the early periods, while concentration of the mining technology creates a further upwards pressure on prices through initial excess demand. It is conceivable that demand would be stronger in markets with asymmetric endowment given the results of Tucker and Xu (2020). Traders who initially do not own the asset might be eager to buy at early on. This might explain how Bitcoin surged from a few cents to $1. In the early days only few computer science aficionados had access to the mining technology (and the knowledge to run it). Potential investors who envisioned Bitcoin to be more valuable in the future but had no access to mining facilities could only buy from the market and thus put upward pressure on prices. Our experiment shows that both the mining protocol as well as the exclusive access to mining technology might be crucial to understand the magnitude and duration of cryptocurrency bubbles.

There are limitations that are important to note regarding our design compared to real world cryptocurrency markets. Firstly, the fundamental value of Bitcoin is a much-debated issue. We concede that our implementation in this aspect is potentially an over-simplification of this. However, the focus of our study is on the effects of mining protocols on overpricing. We conjecture that the effects of mining we identify would persist in situations where the fundamental value would be uncertain and skewed. This would be an interesting avenue left for future research.

Naturally, many other variables could influence price trajectories. In future research, it would be interesting to study other such characteristics of cryptocurrency markets. For example, ambiguity and asymmetric information, especially since ambiguity has been found to be relevant in financial decision making (e.g. Chen and Epstein, 2002; Ju and Miao, 2012). Füllbrunn et al. (2014) do not find effects in market experiments comparing ambiguity and risk. In a recent study, Corgnet et al. (2020) observe that bubbles are less pronounced and do not crash when assets are ambiguous. Oechssler et al. (2011) look into markets with asymmetric information and find that the mere possibility that some traders are better informed than others can create bubbles. It is conceivable that traders believe that this is a possibility in cryptocurrency markets.

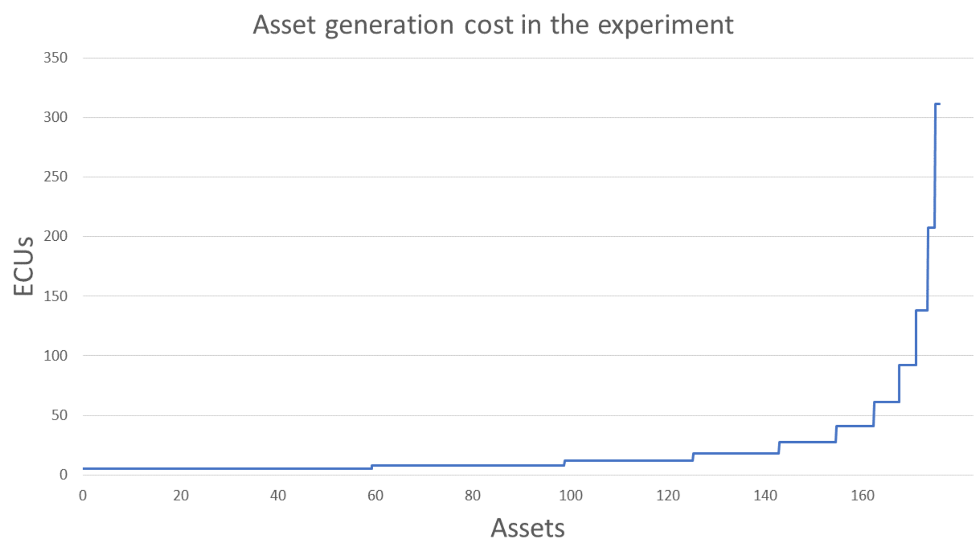
Overall, the observation that mining and concentration of access to the mining technology in a controlled environment fuel overpricing is a highly important result. Any effort put into mining of cryptocurrencies is by design inefficient (see Schilling and Uhlig (2019) for a detailed argument). Furthermore, Auer (2019) explores what the future might hold for cryptocurrencies and concludes that limitations of proof-of-work will ultimately slow down transactions significantly. Similarly, Huberman et al. (2017) and Easley et al. (2019) point out potential inefficiencies and instabilities due to mining.

In a broader picture, our results can potentially inform economists and policy makers in their efforts to develop stable alternative cryptocurrencies. These efforts necessitate further investigation of other mechanisms at play in cryptocurrency markets, which implies there is still plenty of potential for future work in this area.

4.5 Supplementary Material

4.5.1 Asset costs as a function of assets

Figure 4.8: Asset costs as a function of assets in our mining treatments (assuming maximum mining).



4.5.2 Average Prices

Figure 4.9: Weighted average price per period of all sessions in every treatment

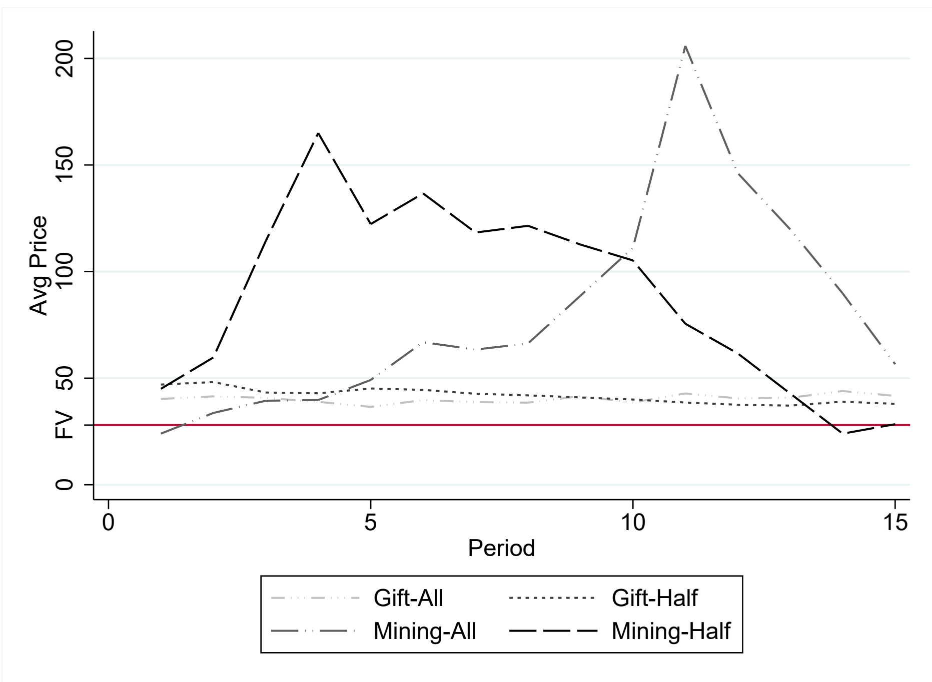
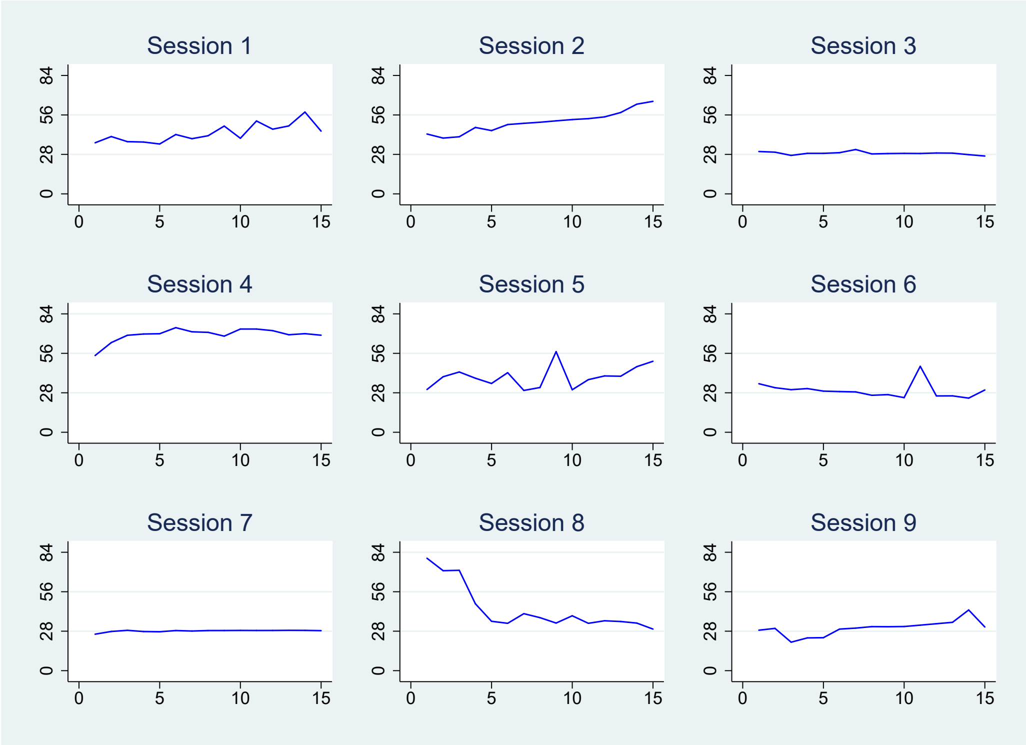


Figure 4.10: Average Prices per period in the individual markets of treatment Gift-All



4.5.3 Individual Sessions

Table 4.8: Bubble measures for the markets in treatment Gift-All

Session	RAD	RD	RDMAX	AMP	CRASH	TURN	LQ	SR	SPREAD	VOLA
1	0.53	0.53	1.07	0.81	-0.48	0.21	1.09	16.52	0.23	0.23
2	0.82	0.82	1.34	0.93	-	0.30	0.19	21.79	0.22	0.14
3	0.04	0.03	0.12	0.15	-0.16	0.12	0.41	18.55	0.05	0.08
4	1.47	1.47	1.65	0.71	-0.22	0.18	0.55	25.55	0.15	0.10
5	0.41	0.41	1.04	0.98	-0.97	0.28	0.20	24.31	0.51	0.46
6	0.13	0.06	0.67	0.79	-0.81	0.12	0.78	19.65	0.11	0.23
7	0.02	0.01	0.02	0.10	-0.01	0.15	0.29	20.88	0.07	0.14
8	0.56	0.56	1.85	-	-1.79	0.27	2.38	24.53	0.79	1.18
9	0.16	0.08	0.54	0.82	-0.43	0.23	0.93	11.72	0.26	0.21

Figures 4.10, 4.11, 4.12 and 4.13 show the individual sessions of each of the treatments. Note that the scale on the y-axis had to be adjusted for about one third of the markets, while the others have a common y-axis, ranging from one to three times the fundamental value. This adjustment was neccessary for only one of the gift markets (i.e. Session 6 in Gift-Half), five markets in treatment Mining-All and six markets in treatment Mining-Half. Price trajectories of treatment Gift-All markets are flat in general. Sessions 1 and 2 show a slight upward tendency over time. Session 8 started on a high price level initially,

Figure 4.11: Average Prices per period in the individual markets of treatment Gift-Half

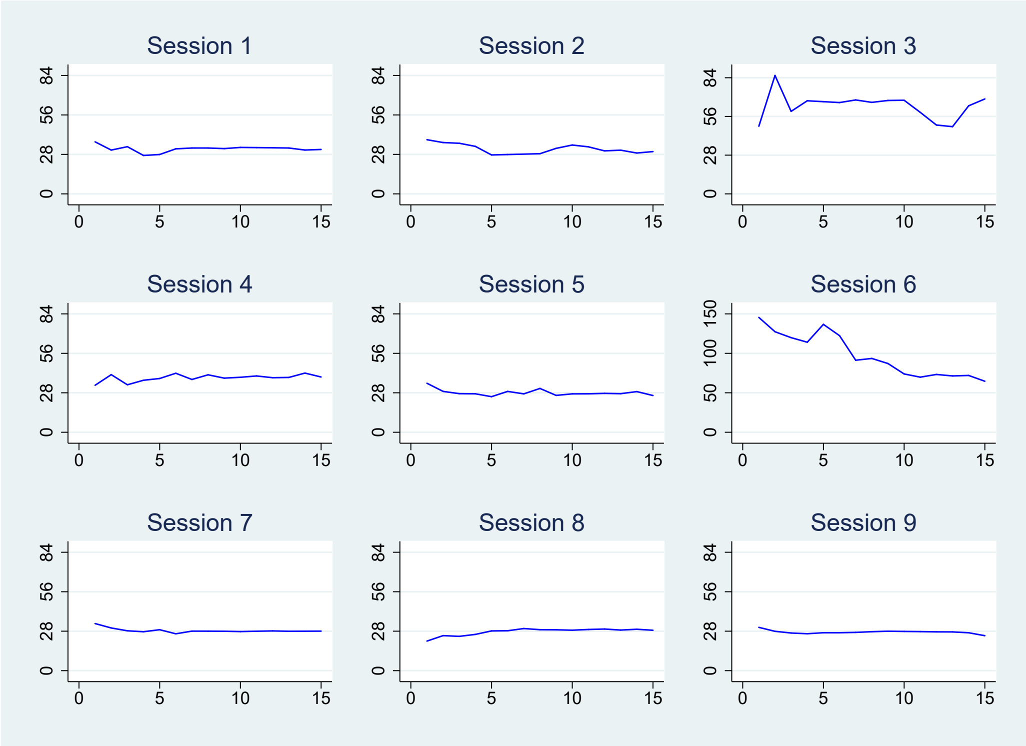


Table 4.9: Bubble measures for the markets in treatment Gift-Half

Session	RAD	RD	RDMAX	AMP	CRASH	TURN	LQ	SR	SPREAD	VOLA
1	0.15	0.14	0.32	-	-0.35	0.29	1.09	19.31	0.10	0.16
2	0.14	0.14	0.37	-	-0.39	0.06	42.46	13.95	0.18	0.05
3	1.27	1.27	2.06	1.31	-1.32	0.14	0.46	18.05	0.31	0.10
4	0.38	0.38	0.50	0.31	-0.10	0.20	0.49	18.27	0.16	0.27
5	0.06	0.00	0.24	-	-0.34	0.20	0.36	19.60	-0.02	0.24
6	2.48	2.48	4.20	-	-2.89	0.21	0.98	27.55	0.88	0.42
7	0.03	0.02	0.19	-	-0.26	0.20	1.72	26.93	0.02	0.04
8	0.06	-0.01	0.07	0.32	-0.04	0.30	1.98	27.33	0.05	0.07
9	0.04	-0.02	0.10	-	-0.21	0.18	0.23	19.02	0.09	0.06

but experienced a downward correction after three periods and stayed flat afterwards. None of the markets seem to have a clear end-of-game effect. The analysis of the price charts of treatment Gift-Half, Figure 4.11, leads to similar conclusions. Most markets have very stable pricing across periods, while session 6 seems to be an exception. In this session, prices started surprisingly high and decreased over time. Again, none of the markets seems to have an end-of-game effect.

The individual markets of treatment Mining-All (Figure 4.12) show a different overall pattern than the gift sessions. Only session 6 shows a flat price trajectory, while all other

Figure 4.12: Average Prices per period in the individual markets of treatment Mining-All

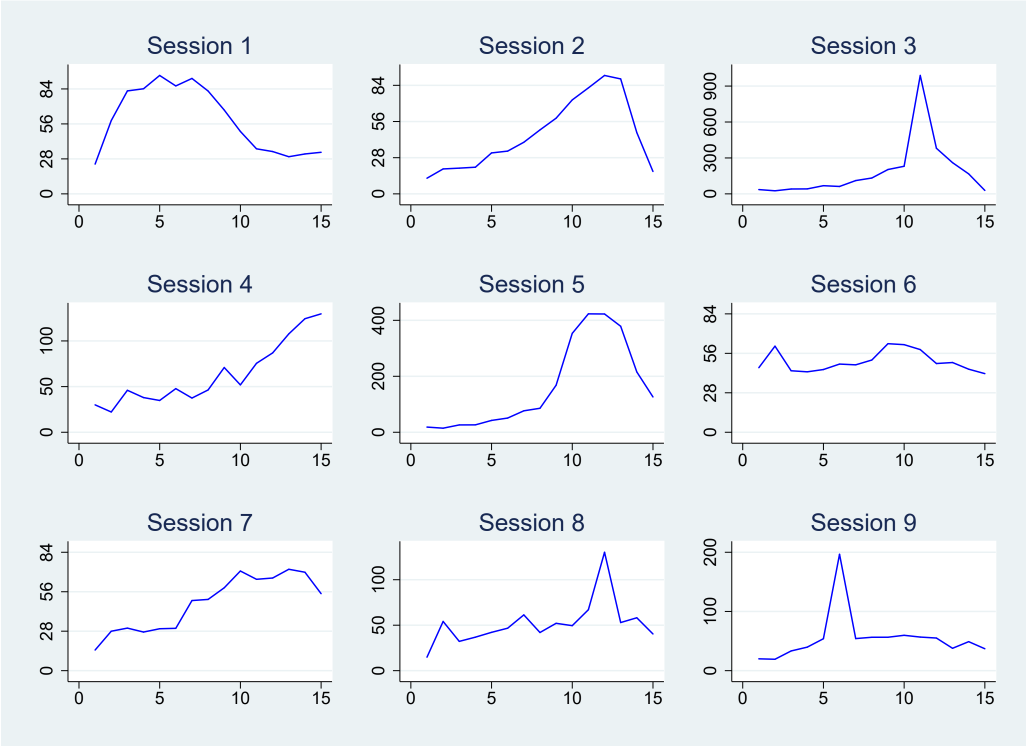


Table 4.10: Bubble measures for the markets in treatment Mining-All

Session	RAD	RD	RDMAX	AMP	CRASH	TURN	LQ	SR	SPREAD	VOLA
1	1.13	1.11	2.39	2.53	-2.32	0.36	2.06	20.72	0.48	0.28
2	0.87	0.63	2.27	2.84	-2.65	0.21	0.54	17.41	0.35	0.29
3	5.62	5.61	34.31	34.40	-34.28	0.18	0.24	12.20	7.57	0.70
4	1.29	1.26	3.63	3.84	-	0.11	0.22	10.18	0.85	0.31
5	4.91	4.78	14.11	14.59	-10.60	0.28	0.64	17.97	1.43	0.30
6	0.79	0.79	1.24	0.71	-0.76	0.25	1.08	19.92	0.32	0.18
7	0.77	0.71	1.57	2.05	-0.62	0.13	0.52	14.26	0.48	0.11
8	0.92	0.86	3.65	4.12	-3.21	0.17	0.43	17.13	0.84	0.61
9	1.05	0.97	6.02	6.33	-5.70	0.14	0.47	16.96	0.48	0.28

markets follow an upward trend in the first periods. Session 4 keeps this trend throughout all periods, i.e. the highest price is reached in the last period. The other seven markets reach a peak price (session 1 and session 9 do so in early periods, sessions 2,3,5,7 and 8 in later periods) and afterwards experience a drop of prices towards the fundamental value of the asset. The magnitude of those peaks and drops differs from market to market. In Figure 4.13 of treatment Mining-Half most markets show a similar trajectory, but again the magnitude differs quite notably. It is noteworthy that most markets reach their peak price in the earlier periods - none of the sessions had their peak price after period 10.

Figure 4.13: Average Prices per period in the individual markets of treatment Mining-Half

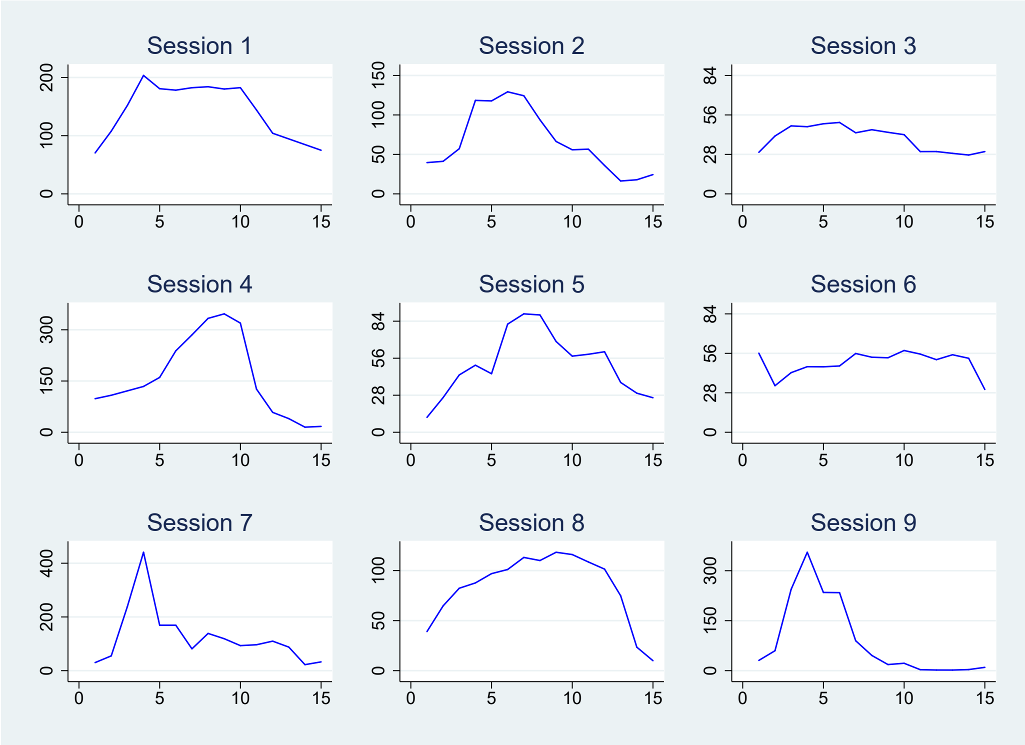


Table 4.11: Bubble measures for the markets in treatment Mining-Half

Session	RAD	RD	RDMAX	AMP	CRASH	TURN	LQ	SR	SPREAD	VOLA
1	3.76	3.76	6.27	4.76	-4.59	0.11	0.30	22.11	1.21	0.22
2	1.49	1.37	3.62	3.20	-4.03	0.31	0.84	25.05	1.76	0.60
3	0.40	0.40	0.81	0.76	-0.83	0.20	1.93	17.37	0.34	0.12
4	4.83	4.72	11.38	8.87	-11.84	0.24	39.40	22.36	1.02	0.08
5	0.94	0.85	2.20	2.80	-2.27	0.34	1.22	20.36	0.96	0.65
6	0.75	0.75	1.07	0.89	-0.99	0.20	0.85	17.86	0.66	0.35
7	3.52	3.49	14.74	14.67	-14.94	0.20	0.40	19.49	1.93	0.87
8	2.08	1.97	3.23	2.83	-3.87	0.33	0.37	23.92	1.32	0.44
9	2.87	2.23	11.69	11.59	-12.62	0.14	1.22	28.16	4.28	0.93

Tables 4.8-4.11 show the bubble measures of each of our treatments. We analyze indicators that were used by Weitzel et al. (2018). Specifically, these indicators include LIQUIDITY. LIQUIDITY describes the volume quantities of open orders at the end of each period. As one can clearly see in Tables 4.9 and 4.11, session 2 in Gift-Half and session 4 in Mining-Half have a puzzling high LIQUIDITY value compared to the other sessions. The interpretation of those values is questionable, as they are based on rather meaningless orders.[66]

[66]For example, in session 4 in treatment Mining-Half, one trader offered to buy 100000 assets for a price of 0.01 ECU each.

4.5.4 Bubble Measures

This section provides the formulas to calculate the bubble measures we use for our analysis. To fix notation, denote:

- T the total number of periods,
- FV_t the fundamental value in period t,
- N_t the total number of trades in period t,
- t^* the period with the highest volume-weighted mean price,
- $\overline{P}_t$ the volume-weighted mean price in period t,
- LO_t the number of shares offered to trade in period t,
- MO_t the number of shares traded based on accepted orders posted by other subjects in period t,
- $R_{t,j}$ the log-return of a trade, i.e. $R_{t,j} = \ln(P_{t,j}/P_{t,j-1})$,
- $\overline{R}_{t,j}$ the average log-return in period t,
- $S_{\hat{t},j}$ the price of sell order j at the end of period t,
- $B_{\hat{t},j}$ the price of buy order j at the end of period t,
- $O_{\hat{t}}$ the number of open orders at the end of period t,
- O_o^j the quantity offered in order o.

Now, define the following bubble measures:

$$RAD = \sum_{t=1}^{T} \frac{\mid \frac{\overline{P}_t - FV_t}{FV_t} \mid}{T}$$

$$RD = \sum_{t=1}^{T} \frac{\frac{\overline{P}_t - FV_t}{FV_t}}{T}$$

$$RDMAX = \max_t \left\{ \frac{\overline{P}_t - FV_t}{FV_t} \right\} = \frac{\overline{P}_{t^*} - FV_{t^*}}{FV_{t^*}}$$

$$AMPLITUDE = \frac{\overline{P}_{t^*} - FV_{t^*}}{FV_{t^*}} - \min_{0 \leq k < t^*} \left\{ \frac{\overline{P}_{t^*-k} - FV_{t^*-k}}{FV_{t^*-k}} \right\}$$

$$CRASH = \min_{0 \leq l \leq T-t^*} \left\{ \frac{\overline{P}_{t^*+l} - FV_{t^*+l}}{FV_{t^*+l}} \right\} - \frac{\overline{P}_{t^*} - FV_{t^*}}{FV_{t^*}}$$

$$SPREAD = \sum_{t=1}^{T} \frac{1}{FV_t} \frac{1}{T} \left[\min_{j \in N_t} \left\{ S_{\hat{t},j} \right\} - \max_{j \in N_t} \left\{ B_{\hat{t},j} \right\} \right]$$

$$VOLA = \sum_{t=1}^{T} \frac{1}{T} \sqrt{\frac{1}{N_t} \sum_{j=1}^{N_t} (R_{t,j} - \overline{R}_t)^2}$$

$$TURNOVER = \sum_{t=1}^{T} \frac{1}{T} \frac{VOL_t}{TSO}$$

$$SR = \sum_{t=1}^{T} \sum_{j=1}^{N_t} \frac{1}{T} \frac{LO_{j,t}}{LO_{j,t} + MO_{j,t}}$$

$$LIQUIDITY = \frac{1}{TSO} \sum_{t=1}^{T} \sum_{o=1}^{O_{\hat{t}}} \frac{1}{T} O_o^j$$

4.5.5 Experimental Implementation

We implemented the experiment using ztree (Fischbacher, 2007) and the technical toolbox GIMS (Palan, 2015). Depending on our treatment, we handed our subjects instructions describing the market. We include the instructions for Gift-half and Mining-half below. Note that the instructions for Gift-All and Mining-All are identical to the respective half versions, except for the endowment parameters (which are the same for every participant in our All-treatments, i.e. no roles A and B are implemented).

1. Allgemeine Informationen

Im nächsten Teil des Experiments geht es um einen Markt für Wertpapiere. Bitte lesen Sie sich diese Anleitung aufmerksam durch. Ihre Entscheidungen beeinflussen Ihre Bezahlung am Ende des Experiments. Sie sollten daher sichergehen, die Funktionen der Plattform fehlerfrei zu beherrschen.
Sie werden zuerst drei Übungsrunden durchlaufen, in denen Sie die Funktionen der Benutzeroberfläche erlernen und ausprobieren können. Diese Übungsrunden werden Ihre Bezahlung nicht beeinflussen. Jede der Übungsrunden dauert 120 Sekunden. Anschließend wird es 15 Handelsrunden geben, welche Ihre Bezahlung beeinflussen. Jede dieser Handelsrunden dauert ebenfalls 120 Sekunden. Sie werden die Möglichkeit haben, Wertpapiere auf einem Markt zu kaufen und zu verkaufen. Die Währung in diesem Markt heißt ECU (Experimental Currency Unit). Sämtliche Handel und Erträge sind in ECUs. Am Anfang des Experiments wird per Zufall die Hälfte der Teilnehmer **Rolle A** zugeteilt, während die andere Hälfte **Rolle B** erhält. Teilnehmer mit Rolle A bekommen 5140 ECUs sowie 40 Einheiten des Wertpapieres. Teilnehmer mit Rolle B erhalten 6260 ECUs sowie 0 Einheiten des Wertpapieres. Alle Teilnehmer können Ihre ECUs benutzen um Wertpapiere auf dem Markt zu kaufen oder zu verkaufen. Ihr Kontostand und Ihr Besitz an Wertpapieren werden von einer Runde in die nächste übertragen.
Am Ende des Experiments wird für alle Teilnehmer per Zufall ermittelt, welchen Wert die eigenen Wertpapiere haben. Hierzu werden 8 Spielkarten benutzt: Zwei Asse, zwei Könige, zwei Damen und zwei Buben. Das Symbol der eigenen Karte entscheidet über den Wert der eigenen Wertpapiere:

Spielkarte	Wert eines Wertpapieres
Ass	67 ECU
König	30 ECU
Dame	15 ECU
Bube	0 ECU

Jeder Teilnehmer zieht der Reihe nach eine Spielkarte, so dass am Ende alle Spielkarten verteilt sind. Dies garantiert, dass genau zwei Teilnehmer ein Ass ziehen, genau zwei Teilnehmer einen König, genau zwei Teilnehmer eine Dame und genau zwei Teilnehmer einen Buben.
Nachdem der Wert Ihrer Wertpapiere ermittelt wurde, werden Sie ausbezahlt. Sie erhalten Euros entsprechend der Summe des ECU-Wertes Ihres Wertpapier-Kontos sowie Ihres ECU-Kontostands. Je mehr ECUs Sie verdienen, desto mehr Euro werden Sie erhalten. Ihre ECUs werden zu folgendem Kurs in Euro umgerechnet:

560 ECUs = 1 Euro

2. Der Markt und Handelsregeln

Marktregeln

Sie können Wertpapiere mit Anderen auf dem Marktplatz handeln. Handel werden in Form von kontinuierlichen Doppel-Auktionen durchgeführt. Das heißt, jeder kann Wertpapiere kaufen und verkaufen.
Wenn Sie einige Einheiten des Wertpapieres kaufen, wird Ihr ECU Kontostand um den fälligen Geldbetrag reduziert (Preis mal Anzahl) und Ihr Bestand an Wertpapieren wird um die gekaufte Anzahl steigen. Wenn Sie Wertpapiere verkaufen, steigt Ihr ECU Kontostand um den entsprechenden Geldbetrag (Preis mal Anzahl) und Ihr Bestand an Wertpapieren wird um die verkaufte Anzahl sinken.
Bitte beachten Sie, dass Sie höchstens so viele Wertpapiere kaufen bzw. verkaufen können, wie durch Ihr Konto gedeckt sind.

Während des Experiments werden Sie einen Bildschirm wie den folgenden sehen:

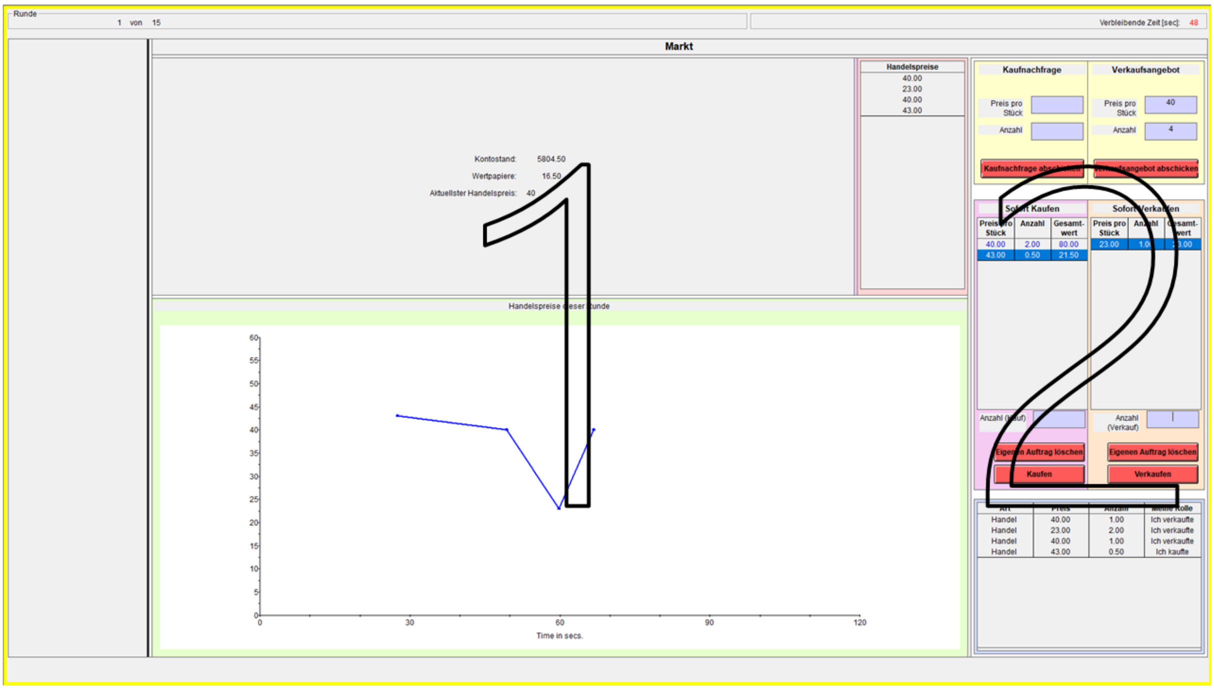


Abbildung 1: Bildschirm

In der Mitte (1) des Bildschirms (siehe Abbildung 1) sehen Sie Informationen zu Ihrem aktuellen Kontostand und Ihren Wertpapieren, sowie eine Preisliste für die aktuelle Handelsrunde. Wenn ein neuer Handel stattfindet wird diese Information in der Liste „Handelspreise" und als neuer Marker in der Preisgrafik unten erscheinen.

Im rechten Segment (2) des Bildschirms (siehe Abbildung 1) finden Sie eine Benutzeroberfläche in der Sie Wertpapiere mit Anderen handeln können.

Marktplatz

Wenn Sie Wertpapiere erwerben möchten können Sie dies auf zwei Wegen tun:

1. Sie können eine **Kaufnachfrage** in dem Kasten „Kaufnachfrage" erstellen, welche dann von einem anderen Teilnehmer akzeptiert werden kann, der an Sie verkaufen möchte. Geben Sie dazu den Preis, den Sie bereit sind für eine Einheit des Wertpapieres zu bezahlen, in das Feld „Preis pro Stück" ein. Geben Sie außerdem die Anzahl der Wertpapiere, die Sie zu diesem Preis kaufen möchten, in das Feld „Anzahl" ein (dies kann auch ein Bruchteil einer Einheit sein). Sie können Ihre Kaufnachfrage einreichen indem Sie auf „Kaufnachfrage abschicken" klicken.
2. Sie können **sofort kaufen**, indem Sie ein Verkaufsangebot aus der Liste im Kasten „Sofort Kaufen" auswählen und die Anzahl, die Sie zu dem angegebenen Preis kaufen möchten, in das Feld „Anzahl (Kauf)" eingeben und dann auf „Kaufen" drücken. Die Liste zeigt alle Verkaufsangebote sortiert nach ihrem Preis, so dass der niedrigste Preis oben steht.

Wenn Sie Wertpapiere verkaufen möchten, haben Sie auch zwei Möglichkeiten:

1. Sie können ein **Verkaufsangebot** in dem Kasten „Verkaufsangebot" erstellen, welcher dann von einem anderen Teilnehmer akzeptiert werden kann, der von Ihnen kaufen möchte. Geben Sie dazu den Preis, zu dem Sie bereit sind eine Einheit des Wertpapieres zu verkaufen, in das

Feld „Preis pro Stück“ ein. Geben Sie außerdem die Anzahl der Wertpapiere, die Sie zu diesem Preis verkaufen möchten, in das Feld „Anzahl“ ein (dies kann auch ein Bruchteil einer Einheit sein). Sie können das Verkaufsangebot einreichen indem Sie auf „Verkaufsangebot abschicken“ klicken.

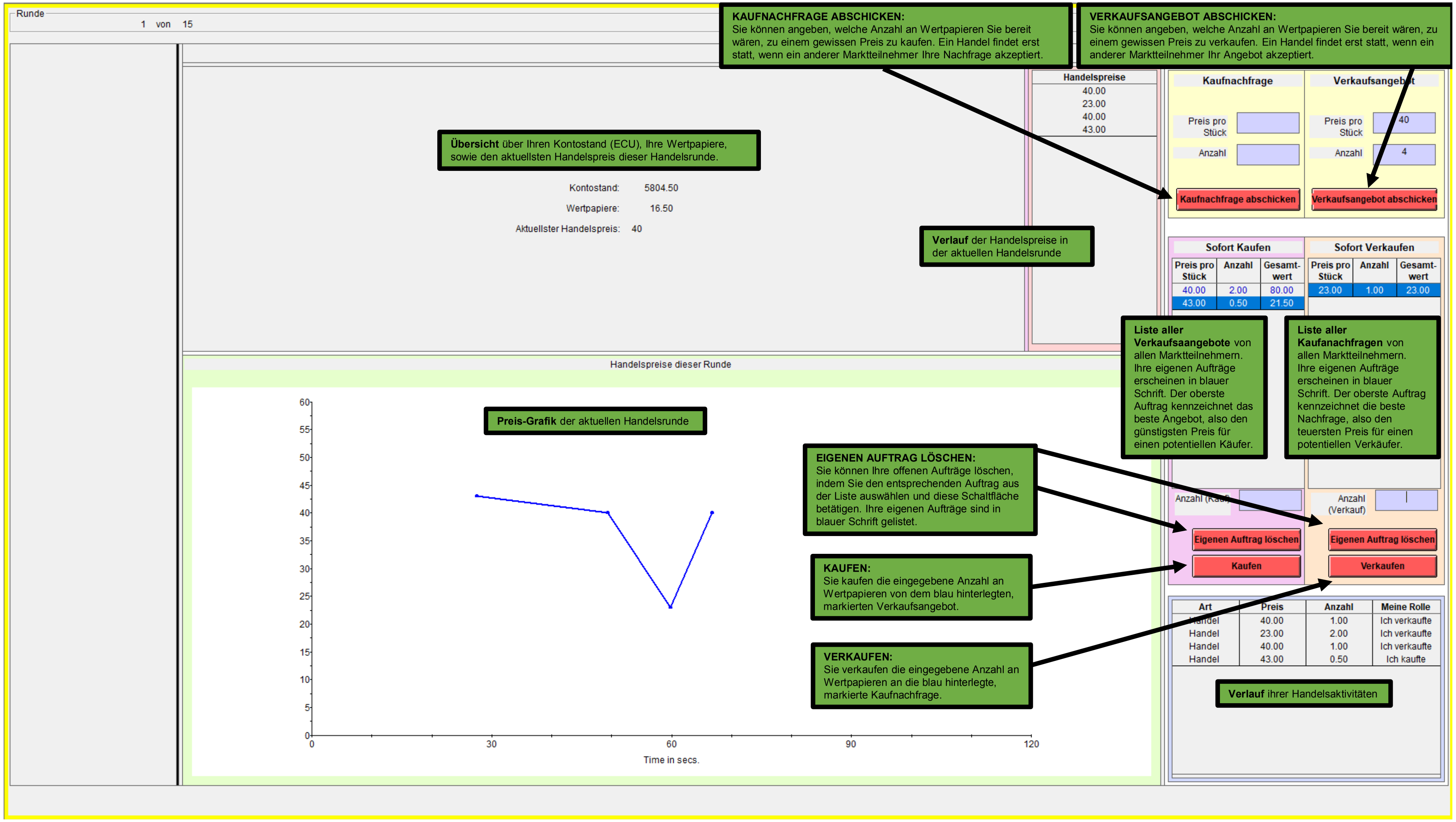
2. Sie können **sofort verkaufen,** indem Sie eine Kaufnachfrage aus der Liste im Kasten „Sofort Verkaufen“ auswählen und die Anzahl, die Sie zu dem angegebenen Preis verkaufen möchten in das Feld „Anzahl (Verkauf)“ eingeben und dann auf „Verkaufen“ klicken. Die Liste zeigt alle Kaufnachfragen sortiert nach ihrem Preis, so dass der höchste Preis oben steht.

Sie können Ihre Kaufnachfragen und Verkaufsangebote zurückziehen, solange Sie nicht von einem anderen Marktteilnehmer akzeptiert wurden. Wählen Sie dafür die entsprechende Zeile in der Liste aus und klicken Sie anschließend auf „Eigene Nachfrage löschen“ bzw. „Eigenes Angebot löschen“. Sie können nur Nachfragen und Angebote löschen, die Sie selbst eingereicht haben. Sie können an der Farbe in der Liste erkennen, welche Aufträge von Ihnen sind. Ihre eigenen Aufträge werden in blauer Schrift sein, die von Anderen in schwarzer Schrift.
Unten im rechten Bereich (2) des Bildschirms sehen Sie eine Liste mit allen Aktionen, an denen Sie beteiligt waren. Wird diese Historie größer als die Tabelle, haben Sie die Möglichkeit zu scrollen, damit Sie die gesamte Historie durchsuchen können.

Am Ende jeder Runde wird ein Bildschirm mit einer Zusammenfassung gezeigt, auf dem Ihr aktueller ECU-Kontostand und Wertpapierbestand angezeigt werden. Außerdem finden Sie dort eine Grafik und eine Liste mit den durchschnittlichen Handelspreisen der vergangenen Runden.

Zusammenfassung:

- Startkapital Rolle A: 5140 ECU, 40 Wertpapiere
- Startkapital Rolle B: 6260 ECU, 0 Wertpapiere
- 3 Übungsrunden à 120 Sekunden
- 15 Handelsrunden à 120 Sekunden
- Kontostände werden von Runde zu Runde übertragen
- Funktionen:
    - Kaufnachfrage
    - Sofort Kaufen
    - Verkaufsangebot
    - Sofort Verkaufen
- Eigene Aufträge in blauer Schrift, fremde Aufträge in schwarzer Schrift
- Am Ende des Marktes:
    - Wertpapiere = 0/15/30/67 ECU
    - 560 ECU = 1 EUR


Runde
1 von 15
KAUFNACHFRAGE ABSCHICKEN:
Sie können angeben, welche Anzahl an Wertpapieren Sie bereit wären, zu einem gewissen Preis zu kaufen. Ein Handel findet erst statt, wenn ein anderer Marktteilnehmer Ihre Nachfrage akzeptiert.
VERKAUFSANGEBOT ABSCHICKEN:
Sie können angeben, welche Anzahl an Wertpapieren Sie bereit wären, zu einem gewissen Preis zu verkaufen. Ein Handel findet erst statt, wenn ein anderer Marktteilnehmer Ihr Angebot akzeptiert.
Handelspreise
40.00
23.00
40.00
43.00
Kaufnachfrage
Verkaufsangebot
Preis pro Stück
Anzahl
40
4
Kaufnachfrage abschicken
Verkaufsangebot abschicken
Übersicht über Ihren Kontostand (ECU), Ihre Wertpapiere, sowie den aktuellsten Handelspreis dieser Handelsrunde.
Kontostand: 5804.50
Wertpapiere: 16.50
Aktuellster Handelspreis: 40
Verlauf der Handelspreise in der aktuellen Handelsrunde
Sofort Kaufen
Sofort Verkaufen
Preis pro Stück | Anzahl | Gesamt-wert
40.00 | 2.00 | 80.00
43.00 | 0.50 | 21.50
23.00 | 1.00 | 23.00
Liste aller Verkaufsangebote von allen Marktteilnehmern. Ihre eigenen Aufträge erscheinen in blauer Schrift. Der oberste Auftrag kennzeichnet das beste Angebot, also den günstigsten Preis für einen potentiellen Käufer.
Liste aller Kaufanachfragen von allen Marktteilnehmern. Ihre eigenen Aufträge erscheinen in blauer Schrift. Der oberste Auftrag kennzeichnet die beste Nachfrage, also den teuersten Preis für einen potentiellen Verkäufer.
Handelspreise dieser Runde
Preis-Grafik der aktuellen Handelsrunde
EIGENEN AUFTRAG LÖSCHEN:
Sie können Ihre offenen Aufträge löschen, indem Sie den entsprechenden Auftrag aus der Liste auswählen und diese Schaltfläche betätigen. Ihre eigenen Aufträge sind in blauer Schrift gelistet.
Anzahl (Verkauf)
Eigenen Auftrag löschen
Kaufen
Verkaufen
KAUFEN:
Sie kaufen die eingegebene Anzahl an Wertpapieren von dem blau hinterlegten, markierten Verkaufsangebot.
VERKAUFEN:
Sie verkaufen die eingegebene Anzahl an Wertpapieren an die blau hinterlegte, markierte Kaufnachfrage.
Art | Preis | Anzahl | Meine Rolle
Handel | 40.00 | 1.00 | Ich verkaufte
Handel | 23.00 | 2.00 | Ich verkaufte
Handel | 40.00 | 1.00 | Ich verkaufte
Handel | 43.00 | 0.50 | Ich kaufte
Verlauf Ihrer Handelsaktivitäten
0 5 10 15 20 25 30 35 40 45 50 55 60
0 30 60 90 120
Time in secs.

1. Allgemeine Informationen

Im nächsten Teil des Experiments geht es um einen Markt für Wertpapiere. Bitte lesen Sie sich diese Anleitung aufmerksam durch. Ihre Entscheidungen beeinflussen Ihre Bezahlung am Ende des Experiments. Sie sollten daher sichergehen, die Funktionen der Plattform fehlerfrei zu beherrschen.
Sie werden zuerst drei Übungsrunden durchlaufen, in denen Sie die Funktionen der Benutzeroberfläche erlernen und ausprobieren können. Diese Übungsrunden werden Ihre Bezahlung nicht beeinflussen. Jede der Übungsrunden dauert 120 Sekunden. Anschließend wird es 15 Handelsrunden geben, welche Ihre Bezahlung beeinflussen. Jede dieser Handelsrunden dauert ebenfalls 120 Sekunden. Sie werden die Möglichkeit haben, Wertpapiere auf einem Markt zu kaufen und zu verkaufen. Die Währung in diesem Markt heißt ECU (Experimental Currency Unit). Sämtliche Handel und Erträge sind in ECUs. Am Anfang des Experiments wird per Zufall die Hälfte der Teilnehmer **Rolle A** zugeteilt, während die andere Hälfte **Rolle B** erhält. Teilnehmer mit Rolle A bekommen 5540 ECUs sowie 0 Einheiten des Wertpapieres, und die Möglichkeit, Wertpapiere zu generieren. Teilnehmer mit Rolle B erhalten 6260 ECUs sowie 0 Einheiten des Wertpapieres und haben keine Möglichkeit, Wertpapiere zu generieren. Alle Teilnehmer können Ihre ECUs benutzen um Wertpapiere auf dem Markt zu kaufen oder zu verkaufen. Wie Teilnehmer mit Rolle A Wertpapiere generieren können wird in einem späteren Abschnitt dieser Anleitung erklärt. Ihr Kontostand und Ihr Besitz an Wertpapieren werden von einer Runde in die nächste übertragen.
Am Ende des Experiments wird für alle Teilnehmer per Zufall ermittelt, welchen Wert die eigenen Wertpapiere haben. Hierzu werden 8 Spielkarten benutzt: Zwei Asse, zwei Könige, zwei Damen und zwei Buben. Das Symbol der eigenen Karte entscheidet über den Wert der eigenen Wertpapiere:

Spielkarte	Wert eines Wertpapieres
Ass	67 ECU
König	30 ECU
Dame	15 ECU
Bube	0 ECU

Jeder Teilnehmer zieht der Reihe nach eine Spielkarte, so dass am Ende alle Spielkarten verteilt sind. Dies garantiert, dass genau zwei Teilnehmer ein Ass ziehen, genau zwei Teilnehmer einen König, genau zwei Teilnehmer eine Dame und genau zwei Teilnehmer einen Buben.
Nachdem der Wert Ihrer Wertpapiere ermittelt wurde, werden Sie ausbezahlt. Sie erhalten Euros entsprechend der Summe des ECU-Wertes Ihres Wertpapier-Kontos sowie Ihres ECU-Kontostands. Je mehr ECUs Sie verdienen, desto mehr Euro werden Sie erhalten. Ihre ECUs werden zu folgendem Kurs in Euro umgerechnet:

560 ECUs = 1 Euro

2. Generieren von Wertpapieren, der Markt und Handelsregeln

Marktregeln

Sie können Wertpapiere mit Anderen auf dem Marktplatz handeln. Handel werden in Form von kontinuierlichen Doppel-Auktionen durchgeführt. Das heißt, jeder kann Wertpapiere kaufen und verkaufen.
Wenn Sie einige Einheiten des Wertpapiers kaufen, wird Ihr ECU Kontostand um den fälligen Geldbetrag reduziert (Preis mal Anzahl) und Ihr Bestand an Wertpapieren wird um die gekaufte Anzahl steigen. Wenn Sie Wertpapiere verkaufen, steigt Ihr ECU Kontostand um den entsprechenden Geldbetrag (Preis mal Anzahl) und Ihr Bestand an Wertpapieren wird um die verkaufte Anzahl sinken.

Bitte beachten Sie, dass Sie höchstens so viele Wertpapiere kaufen bzw. verkaufen können, wie durch Ihr Konto gedeckt sind.

Während des Experiments werden Sie einen Bildschirm wie den folgenden sehen:

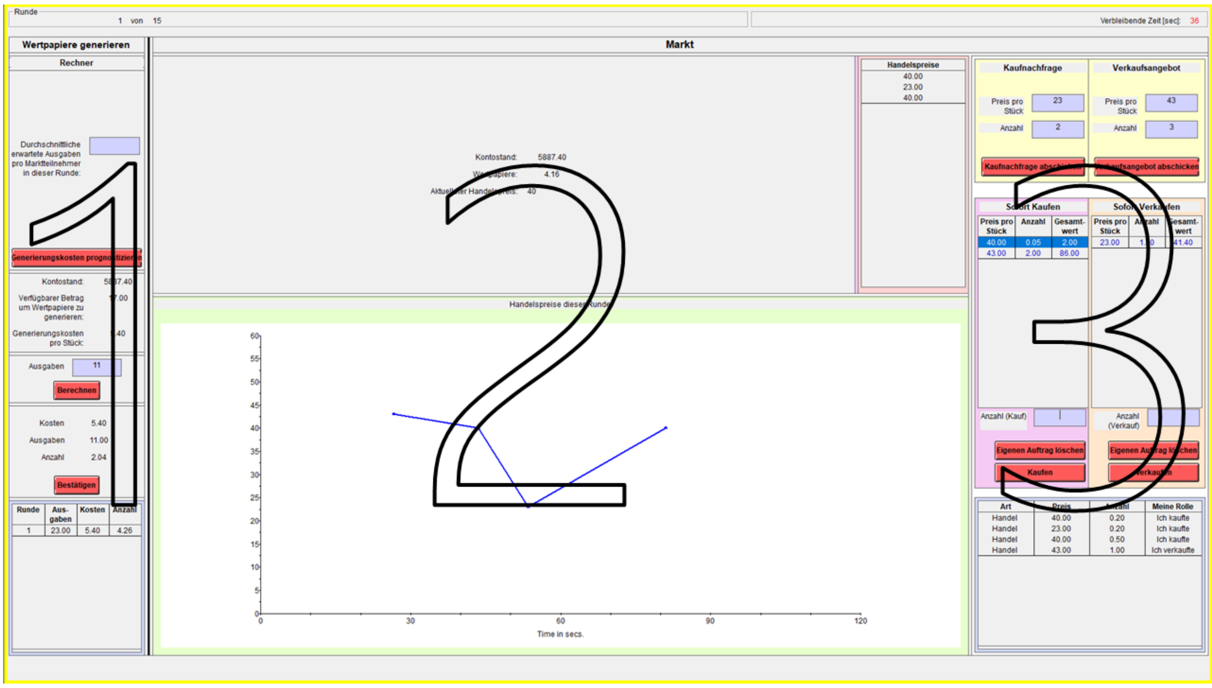


Abbildung 1: Bildschirm

Der Bildschirm ist in verschiedene Segmente unterteilt (siehe Abbildung 1). Das linke Segment (1) dient der Generierung von Wertpapieren. In der Mitte (2) des Bildschirms sehen Sie Informationen zu Ihrem aktuellen Kontostand und Ihren Wertpapieren, sowie eine Preisliste für die aktuelle Handelsrunde. Wenn ein neuer Handel stattfindet wird diese Information in der Liste „Handelspreise“ und als neuer Marker in der Preisgrafik unten erscheinen.

Im rechten Segment (3) des Bildschirms finden Sie eine Benutzeroberfläche in der Sie Wertpapiere mit Anderen handeln können.

Der folgende Abschnitt erklärt zunächst wie Sie ein Wertpapier generieren können. Anschließend werden die Funktionen des Marktplatzes beschrieben.

Generieren von Wertpapieren

Im linken Bereich (1) können Sie in jeder Handelsrunde entscheiden, ob Sie einige Ihrer ECUs zur Generierung von Wertpapieren ausgeben möchten. Beachten Sie, dass Sie in jeder Runde höchstens 80 ECUs zur Generierung von Wertpapieren ausgeben können, sofern Ihnen Rolle A zugeteilt wurde. Falls Ihnen Rolle B zugeteilt wird können sie 0 ECUs zur Generierung von Wertpapieren ausgeben. Die Kosten um Wertpapiere zu generieren variieren über die Zeit. Die Kosten bleiben in jeder Runde konstant, werden aber zu Beginn der folgenden Runde neu berechnet. Die Kosten der Generierung hängen davon ab, wie viele ECUs in allen vorhergehenden Runden insgesamt von allen Marktteilnehmern zur Generierung ausgegeben wurden. Abbildung 2 zeigt wie die Kosten von den Gesamtausgaben für die Generierung von Wertpapieren abhängen. Die vertikale Achse zeigt die Generierungskosten pro Wertpapier, die horizontale Achse zeigt die Gesamtausgaben (alle Ausgaben über alle vorherigen Runden aller Teilnehmer*innen aufaddiert). Beachten Sie, dass die Kosten der Generierung nur steigen können, sie werden nie sinken.

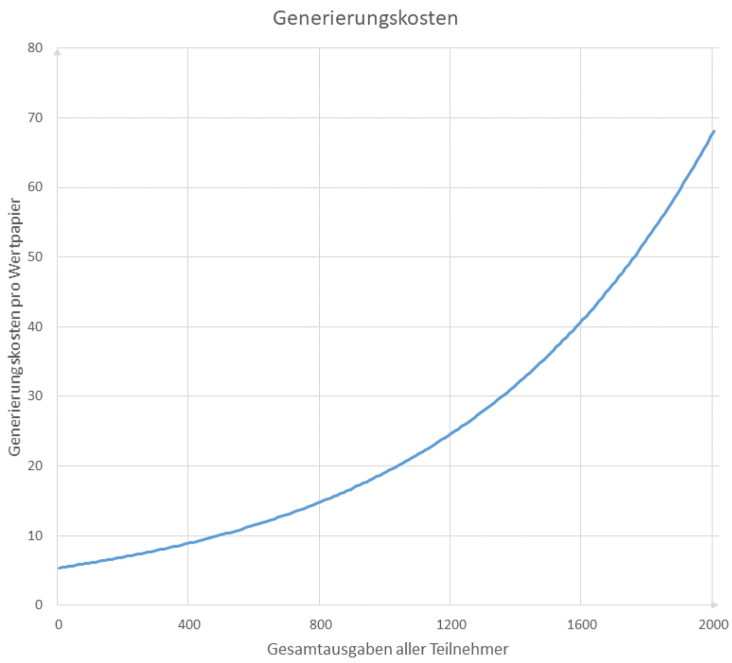


Abbildung 2: Kosten der Generierung von Wertpapieren

Der Bildschirm für die Generierung von Wertpapieren (Segment 1) besteht aus drei Teilen. Ganz oben befindet sich ein Rechner der Ihnen dabei hilft, die Kosten der Generierung in den folgenden Runden zu berechnen. In dem Feld „Durchschnittliche erwartete Ausgaben pro Marktteilnehmer in dieser Runde“ können Sie eine Zahl eingeben, von der Sie denken, dass die Teilnehmer*innen diese in der aktuellen Runde durchschnittlich ausgeben. Wenn Sie auf „Generierungskosten prognostizieren“ klicken, wird eine Tabelle erscheinen, die anzeigt, wie sich die Kosten zur Generierung in den nächsten vier Runden entwickeln werden (unter der Annahme, dass die Anderen in jeder Runde so viel ausgeben, wie Sie angegeben haben). In der Mitte des linken Segments (1) können Sie Wertpapiere generieren. Sie finden dort Informationen über die ECUs, über die Sie insgesamt verfügen, sowie die begrenzte Anzahl der ECUs, die Ihnen zur Generierung von Wertpapieren zur Verfügung stehen (dieser Wert wird zu Beginn jeder Runde auf 80 ECUs zurückgesetzt, sofern Ihnen Rolle A zugeteilt wurde). Außerdem finden Sie dort die aktuellen Kosten der Generierung eines Wertpapiers. Zu Beginn jeder neuen Runde werden die Kosten wie in der Abbildung oben dargestellt berechnet. Die Kosten beziehen sich immer auf genau ein Wertpapier. Es ist aber ebenfalls möglich, Teile eines Wertpapiers zu generieren. Zur Generierung geben Sie die Menge an ECUs, die Sie ausgeben möchten, in das Feld „Ausgaben“ ein. Wenn Sie dann auf den Button „Berechnen“ klicken, sehen Sie wie viele Wertpapiere Sie mit diesen Ausgaben generieren können. Wenn Sie die Generierung fortsetzen wollen, können Sie dies durch Klicken auf „Bestätigen“ tun. Möchten Sie die Höhe der Ausgaben ändern, können Sie einfach die Zahl im Feld „Ausgaben“ ändern und erneut auf „Berechnen“ klicken. Sie sehen ein Beispiel für diese Prozedur in Abbildung 3. Falls Sie Ihre Generierung bestätigen, wird Ihr Kontostand umgehend aktualisiert, d.h. die entsprechenden ECUs werden Ihnen abgezogen, und Ihr Wertpapierbestand wird erhöht.

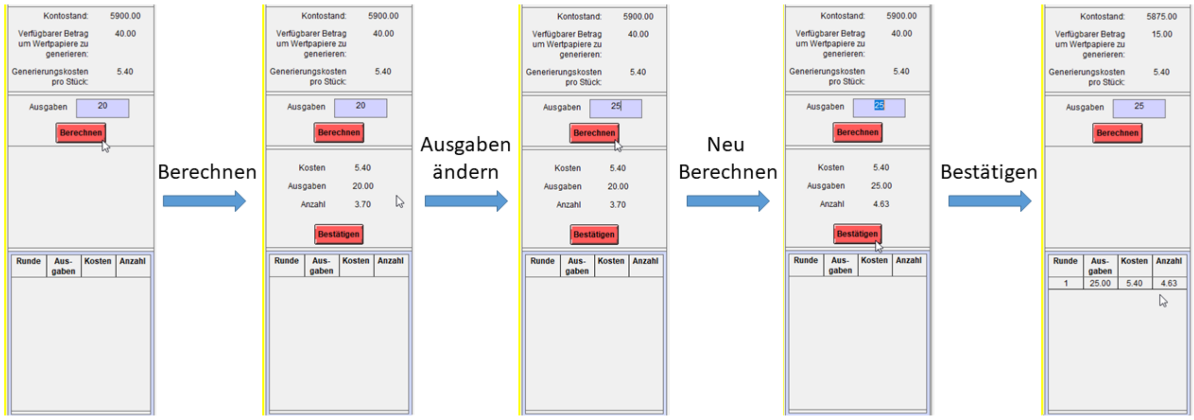


Abbildung 3: Beispiel zur Generierung von Wertpapieren

Im unteren Teil des linken Bereiches (1) wird Ihnen Ihre persönliche Generierungs-Historie aufgelistet. Jede Generierung von Wertpapieren die Sie bestätigen wird hier aufgelistet. Falls Ihre Historie zu groß für den Platz der Tabelle ist können Sie in dieser scrollen.

Marktplatz

Wenn Sie Wertpapiere erwerben möchten können Sie dies auf zwei Wegen tun:

1. Sie können eine **Kaufnachfrage** in dem Kasten „Kaufnachfrage" erstellen, welche dann von einem anderen Teilnehmer akzeptiert werden kann, der an Sie verkaufen möchte. Geben Sie dazu den Preis, den Sie bereit sind für eine Einheit des Wertpapiers zu bezahlen, in das Feld „Preis pro Stück" ein. Geben Sie außerdem die Anzahl der Wertpapiere, die Sie zu diesem Preis kaufen möchten, in das Feld „Anzahl" ein (dies kann auch ein Bruchteil einer Einheit sein). Sie können Ihre Kaufnachfrage einreichen indem Sie auf „Kaufnachfrage abschicken" klicken.
2. Sie können **sofort kaufen**, indem Sie ein Verkaufsangebot aus der Liste im Kasten „Sofort Kaufen" auswählen und die Anzahl, die Sie zu dem angegebenen Preis kaufen möchten, in das Feld „Anzahl (Kauf)" eingeben und dann auf „Kaufen" drücken. Die Liste zeigt alle Verkaufsangebote sortiert nach ihrem Preis, so dass der niedrigste Preis oben steht.

Wenn Sie Wertpapiere verkaufen möchten, haben Sie auch zwei Möglichkeiten:

1. Sie können ein **Verkaufsangebot** in dem Kasten „Verkaufsangebot" erstellen, welcher dann von einem anderen Teilnehmer akzeptiert werden kann, der von Ihnen kaufen möchte. Geben Sie dazu den Preis, zu dem Sie bereit sind eine Einheit des Wertpapiers zu verkaufen, in das Feld „Preis pro Stück" ein. Geben Sie außerdem die Anzahl der Wertpapiere, die Sie zu diesem Preis verkaufen möchten, in das Feld „Anzahl" ein (dies kann auch ein Bruchteil einer Einheit

sein). Sie können das Verkaufsangebot einreichen indem Sie auf „Verkaufsangebot abschicken“ klicken.

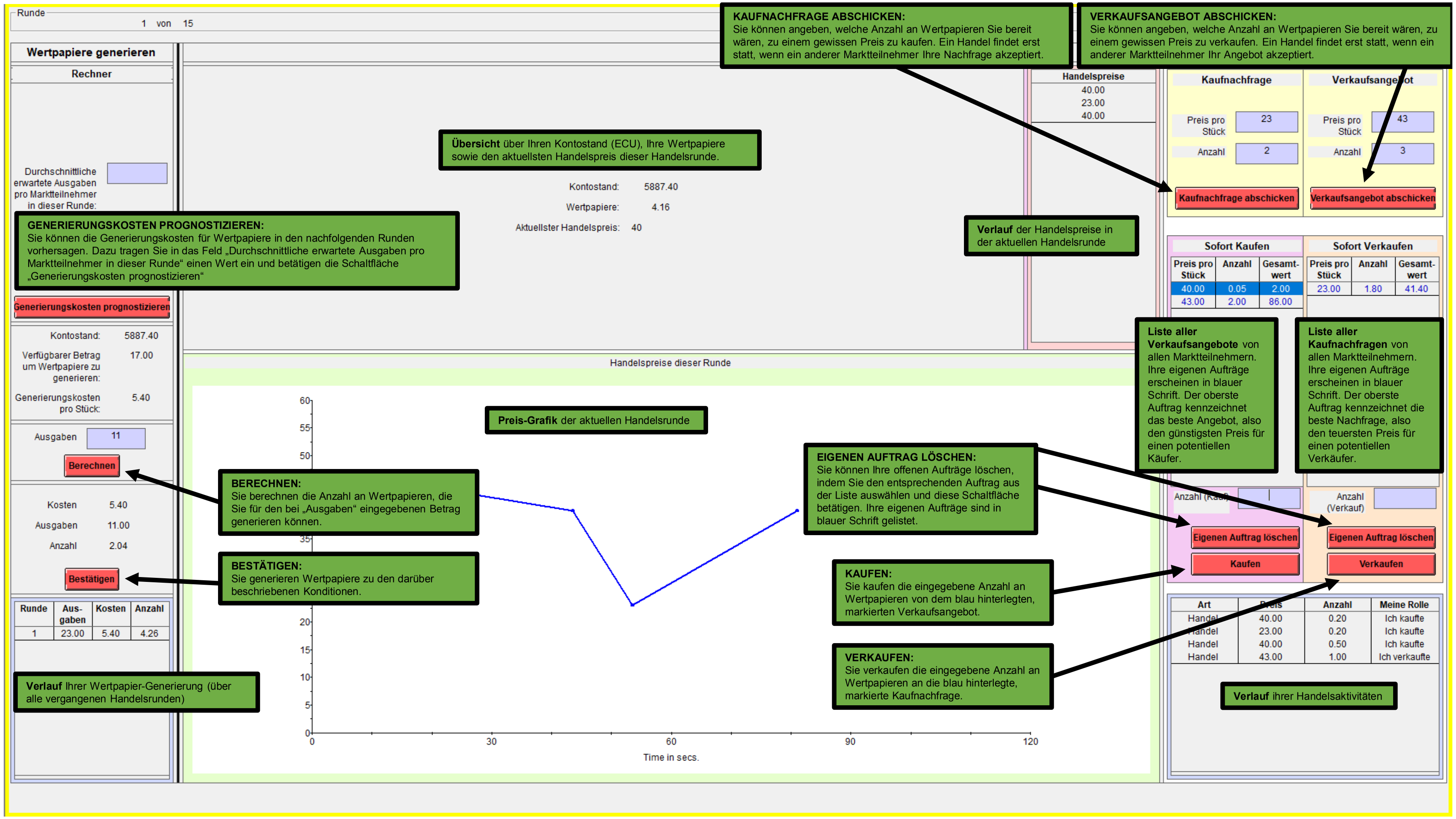
2. Sie können **sofort verkaufen,** indem Sie eine Kaufnachfrage aus der Liste im Kasten „Sofort Verkaufen“ auswählen und die Anzahl, die Sie zu dem angegebenen Preis verkaufen möchten in das Feld „Anzahl (Verkauf)“ eingeben und dann auf „Verkaufen“ klicken. Die Liste zeigt alle Kaufnachfragen sortiert nach ihrem Preis, so dass der höchste Preis oben steht.

Sie können Ihre Kaufnachfragen und Verkaufsangebote zurückziehen, solange Sie nicht von einem anderen Marktteilnehmer akzeptiert wurden. Wählen Sie dafür die entsprechende Zeile in der Liste aus und klicken Sie anschließend auf „Eigene Nachfrage löschen“ bzw. „Eigenes Angebot löschen“. Sie können nur Nachfragen und Angebote löschen, die Sie selbst eingereicht haben. Sie können an der Farbe in der Liste erkennen, welche Aufträge von Ihnen sind. Ihre eigenen Aufträge werden in blauer Schrift sein, die von Anderen in schwarzer Schrift.
Unten im rechten Bereich (2) des Bildschirms sehen Sie eine Liste mit allen Aktionen, an denen Sie beteiligt waren. Wird diese Historie größer als die Tabelle, haben Sie die Möglichkeit zu scrollen, damit Sie die gesamte Historie durchsuchen können.

Am Ende jeder Runde wird ein Bildschirm mit einer Zusammenfassung gezeigt, auf dem Ihr aktueller ECU-Kontostand und Wertpapierbestand sowie Informationen zur Generierung angezeigt werden. Außerdem finden Sie dort eine Grafik und eine Liste mit den durchschnittlichen Handelspreisen der vergangenen Runden.

Zusammenfassung:

- Startkapital Rolle A: 5540 ECU, 0 Wertpapiere
- Startkapital Rolle B: 6260 ECU, 0 Wertpapiere
- 3 Übungsrunden à 120 Sekunden
- 15 Handelsrunden à 120 Sekunden
- Kontostände werden von Runde zu Runde übertragen
- Funktionen:
    - Wertpapiergenerierung
    - Kaufnachfrage
    - Sofort Kaufen
    - Verkaufsangebot
    - Sofort Verkaufen
- Generierungslimit Rolle A: 80 ECU
- Generierungslimit Rolle B: 0 ECU
- Generierungskosten steigen zu Beginn jeder Runde, solange die Gesamtausgaben aller Teilnehmer steigen
- Eigene Aufträge in blauer Schrift, fremde Aufträge in schwarzer Schrift
- Am Ende des Marktes:
    - Wertpapiere = 0/15/30/67 ECU
    - 560 ECU = 1 EUR



Runde
1 von 15
Wertpapiere generieren
Rechner
Durchschnittliche erwartete Ausgaben pro Marktteilnehmer in dieser Runde:
GENERIERUNGSKOSTEN PROGNOSTIZIEREN:
Sie können die Generierungskosten für Wertpapiere in den nachfolgenden Runden vorhersagen. Dazu tragen Sie in das Feld „Durchschnittliche erwartete Ausgaben pro Marktteilnehmer in dieser Runde“ einen Wert ein und betätigen die Schaltfläche „Generierungskosten prognostizieren“
Generierungskosten prognostizieren
Kontostand: 5887.40
Verfügbarer Betrag um Wertpapiere zu generieren: 17.00
Generierungskosten pro Stück: 5.40
Ausgaben 11
Berechnen
BERECHNEN:
Sie berechnen die Anzahl an Wertpapieren, die Sie für den bei „Ausgaben“ eingegebenen Betrag generieren können.
Kosten 5.40
Ausgaben 11.00
Anzahl 2.04
Bestätigen
BESTÄTIGEN:
Sie generieren Wertpapiere zu den darüber beschriebenen Konditionen.
Runde | Aus-gaben | Kosten | Anzahl
1 | 23.00 | 5.40 | 4.26
Verlauf Ihrer Wertpapier-Generierung (über alle vergangenen Handelsrunden)
Übersicht über Ihren Kontostand (ECU), Ihre Wertpapiere sowie den aktuellsten Handelspreis dieser Handelsrunde.
Kontostand: 5887.40
Wertpapiere: 4.16
Aktuellster Handelspreis: 40
KAUFNACHFRAGE ABSCHICKEN:
Sie können angeben, welche Anzahl an Wertpapieren Sie bereit wären, zu einem gewissen Preis zu kaufen. Ein Handel findet erst statt, wenn ein anderer Marktteilnehmer Ihre Nachfrage akzeptiert.
VERKAUFSANGEBOT ABSCHICKEN:
Sie können angeben, welche Anzahl an Wertpapieren Sie bereit wären, zu einem gewissen Preis zu verkaufen. Ein Handel findet erst statt, wenn ein anderer Marktteilnehmer Ihr Angebot akzeptiert.
Handelspreise
40.00
23.00
40.00
Verlauf der Handelspreise in der aktuellen Handelsrunde
Kaufnachfrage
Preis pro Stück 23
Anzahl 2
Kaufnachfrage abschicken
Verkaufsangebot
Preis pro Stück 43
Anzahl 3
Verkaufsangebot abschicken
Sofort Kaufen
Preis pro Stück | Anzahl | Gesamt-wert
40.00 | 0.05 | 2.00
43.00 | 2.00 | 86.00
Sofort Verkaufen
Preis pro Stück | Anzahl | Gesamt-wert
23.00 | 1.80 | 41.40
Liste aller Verkaufsangebote von allen Marktteilnehmern. Ihre eigenen Aufträge erscheinen in blauer Schrift. Der oberste Auftrag kennzeichnet das beste Angebot, also den günstigsten Preis für einen potentiellen Käufer.
Liste aller Kaufnachfragen von allen Marktteilnehmern. Ihre eigenen Aufträge erscheinen in blauer Schrift. Der oberste Auftrag kennzeichnet die beste Nachfrage, also den teuersten Preis für einen potentiellen Verkäufer.
Handelspreise dieser Runde
Preis-Grafik der aktuellen Handelsrunde
Time in secs.
EIGENEN AUFTRAG LÖSCHEN:
Sie können Ihre offenen Aufträge löschen, indem Sie den entsprechenden Auftrag aus der Liste auswählen und diese Schaltfläche betätigen. Ihre eigenen Aufträge sind in blauer Schrift gelistet.
Anzahl (Verkauf)
Eigenen Auftrag löschen
Kaufen
Verkaufen
KAUFEN:
Sie kaufen die eingegebene Anzahl an Wertpapieren von dem blau hinterlegten, markierten Verkaufsangebot.
VERKAUFEN:
Sie verkaufen die eingegebene Anzahl an Wertpapieren an die blau hinterlegte, markierte Kaufnachfrage.
Art | Preis | Anzahl | Meine Rolle
Handel | 40.00 | 0.20 | Ich kaufte
Handel | 23.00 | 0.20 | Ich kaufte
Handel | 40.00 | 0.50 | Ich kaufte
Handel | 43.00 | 1.00 | Ich verkaufte
Verlauf ihrer Handelsaktivitäten

List of Figures

List of Tables

References

ALLISON, P. D. AND N. A. CHRISTAKIS (1994): "Logit models for sets of ranked items," *Sociological methodology*, 199–228.

AUER, R. (2019): "Beyond the doomsday economics of 'proof-of-work' in cryptocurrencies," *BIS Working Papers No. 765.*

BAUR, D. G., K. HONG, AND A. D. LEE (2018): "Bitcoin: Medium of exchange or speculative assets?" *Journal of International Financial Markets, Institutions and Money*, 54, 177–189.

BEGGS, S., S. CARDELL, AND J. HAUSMAN (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., D. HILTON, K. MAZURIER, AND S. POUGET (2005): "Judgemental overconfidence, self-monitoring, and trading performance in an experimental financial market," *The Review of Economic Studies*, 72, 287–312.

BIANCHETTI, M., C. RICCI, AND M. SCARINGI (2018): "Are cryptocurrencies real financial bubbles? Evidence from quantitative analyses," *Evidence from Quantitative Analyses (February 24, 2018). A version of this paper was published in Risk*, 26.

BINDE, P., U. ROMILD, AND R. A. VOLBERG (2017): "Forms of gambling, gambling involvement and problem gambling: evidence from a Swedish population survey," *International Gambling Studies*, 17, 490–507.

Böhme, R., N. Christin, B. Edelman, and T. Moore (2015): "Bitcoin: Economics, technology, and governance," *Journal of Economic Perspectives*, 29, 213–38.

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

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

Bradley, R. A. and M. E. Terry (1952): "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, 39, 324–345.

Breitmoser, Y. (2019): "The axiomatic foundation of logit," *University of Bielefeld.*

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

Bruguier, A. J., S. R. Quartz, and P. Bossaerts (2010): "Exploring the nature of "trader intuition"," *The Journal of Finance*, 65, 1703–1723.

Brunnermeier, M. K. and I. Schnabel (2016): *Bubbles and Central Banks*, Cambridge University Press, 493–562, Studies in Macroeconomic History.

Burniske, C. and A. White (2017): "Bitcoin: Ringing the bell for a new asset class," *Ark Invest (January 2017) https://research. ark-invest. com/hubfs/1_Download_Files_ARK-Invest/White_Papers/Bitcoin-Ringing-The-Bell-For-A-New-Asset-Class. pdf.*

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

Camerer, C. and D. Lovallo (1999): "Overconfidence and excess entry: An experimental approach," *American Economic Review*, 89, 306–318.

Cheah, E.-T. and J. Fry (2015): "Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of Bitcoin," *Economics Letters*, 130, 32–36.

Chen, Z. and L. Epstein (2002): "Ambiguity, risk, and asset returns in continuous time," *Econometrica*, 70, 1403–1443.

Corgnet, B., M. Desantis, and D. Porter (2018): "What makes a good trader? On the role of intuition and reflection on trader performance," *The Journal of Finance*, 73, 1113–1137.

Corgnet, B., R. Hernán-González, and P. Kujal (2020): "On booms that never bust: Ambiguity in experimental asset markets with bubbles," *Journal of Economic Dynamics and Control*, 110, 103754.

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

——— (2008b): "Poker superstars: Skill or luck? Similarities between golf—thought to be a game of skill—and poker," *Chance*, 21, 25–28.

Cueva, C. and A. Rustichini (2015): "Is financial instability male-driven? Gender and cognitive skills in experimental asset markets," *Journal of Economic Behavior & Organization*, 119, 330–344.

Debreu, G. (1960): "Review of 'Individual Choice Behavior' by R. Luce," *The American Economic Review*, 50, 186–188.

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

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

——— (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, 255–264.

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

Dufwenberg, M., R. Sundaram, and D. J. Butler (2010): "Epiphany in the Game of 21," *Journal of Economic Behavior & Organization*, 75, 132–143.

Easley, D., M. O'Hara, and S. Basu (2019): "From mining to markets: The evolution of bitcoin transaction fees," *Journal of Financial Economics*, 134, 91–109.

Eckel, C. C. and S. C. Füllbrunn (2015): "Thar she blows? Gender, competition, and bubbles in experimental asset markets," *American Economic Review*, 105, 906–20.

Eckel, C. C. and P. J. Grossman (2008): "Men, women and risk aversion: Experimental evidence," *Handbook of experimental economics results*, 1, 1061–1073.

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

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

Filippin, A., U. Schmidt, and M. Tomasuolo (2020): "Is gambling more dangerous than betting?" *Unpublished manuscript.*

Fischbacher, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental economics*, 10, 171–178.

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

Füllbrunn, S., H. A. Rau, and U. Weitzel (2014): "Does ambiguity aversion survive in experimental asset markets?" *Journal of Economic Behavior & Organization*, 107, 810–826.

Gervais, A., G. O. Karame, K. Wüst, V. Glykantzis, H. Ritzdorf, and S. Capkun (2016): "On the security and performance of proof of work blockchains," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 3–16.

Glaser, F., K. Zimmermann, M. Haferkorn, M. C. Weber, and M. Siering (2014): "Bitcoin-asset or currency? revealing users' hidden intentions," *Revealing Users' Hidden Intentions (April 15, 2014). ECIS.*

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

Greiner, B. (2015): "Subject pool recruitment procedures: organizing experiments with ORSEE," *Journal of the Economic Science Association*, 1, 114–125.

Gul, F., P. Natenzon, and W. Pesendorfer (2014): "Random choice as behavioral optimization," *Econometrica*, 82, 1873–1912.

Guo, F., C. R. Chen, and Y. S. Huang (2011): "Markets contagion during financial crisis: A regime-switching approach," *International Review of Economics & Finance*, 20, 95–109.

Gurdgiev, C. and D. O'Loughlin (2020): "Herding and anchoring in cryptocurrency markets: Investor reaction to fear and uncertainty," *Journal of Behavioral and Experimental Finance*, 25, 100271.

Hefti, A., S. Heinke, and F. Schneider (2016): "Mental capabilities, trading styles, and asset market bubbles: theory and experiment," Tech. rep.

Heider, F. and M. Simmel (1944): "An experimental study of apparent behavior," *The American journal of psychology*, 57, 243–259.

Holt, C. A., M. Porzio, and M. Y. Song (2017): "Price bubbles, gender, and expectations in experimental asset markets," *European Economic Review*, 100, 72–94.

Holznagel, B. (2008): "Poker - Glücks- oder Geschicklichkeitsspiel," *Multimedia und Recht*, 11, 439–444.

Hong, K. (2017): "Bitcoin as an alternative investment vehicle," *Information Technology and Management*, 18, 265–275.

Huberman, G., J. Leshno, and C. C. Moallemi (2017): "Monopoly without a monopolist: An economic analysis of the bitcoin payment system," *Bank of Finland Research Discussion Paper*.

Ju, N. and J. Miao (2012): "Ambiguity, learning, and asset returns," *Econometrica*, 80, 559–591.

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.

Kirchler, M., J. Huber, and T. Stöckl (2012): "Thar she bursts: Reducing confusion reduces bubbles," *American Economic Review*, 102, 865–83.

Krafft, P. M., N. Della Penna, and A. S. Pentland (2018): "An experimental study of cryptocurrency market dynamics," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13.

Kristoufek, L. (2015): "What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis," *PloS one*, 10, e0123923.

Larkey, P., J. B. Kadane, R. Austin, and S. Zamir (1997): "Skill in games," *Management Science*, 43, 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, 31–44.

Luce, R. D. (1959): *Individual choice behavior: A theoretical analysis*, John Wiley & Sons, Inc., New York.

Malmendier, U. and G. Tate (2005): "CEO overconfidence and corporate investment," *The Journal of Finance*, 60, 2661–2700.

——— (2008): "Who makes acquisitions? CEO overconfidence and the market's reaction," *Journal of Financial Economics*, 89, 20–43.

Manaa, M., M. T. Chimienti, M. M. Adachi, P. Athanassiou, I. Balteanu, A. Calza, C. Devaney, E. Diaz Fernandez, F. Eser, I. Ganoulis, et al. (2019): "Crypto-Assets: Implications for financial stability, monetary policy, and payments and market infrastructures," *ECB Occasional Paper*, No. 223.

Manzini, P. and M. Mariotti (2014): "Stochastic choice and consideration sets," *Econometrica*, 82, 1153–1176.

McFadden, D. (1973): "Conditional logit analysis of qualitative choice behavior," *Frontiers of Econometrics, ed. P. Zarembka. New York: Academic Press*, 105–142.

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

Nakamoto, S. et al. (2008): "Bitcoin: A peer-to-peer electronic cash system," *https://bitcoin.org/bitcoin.pdf*, last accessed on Dec 3rd, 2019.

Noussair, C., S. Robin, and B. Ruffieux (2001): "Price bubbles in laboratory asset markets with constant fundamental values," *Experimental Economics*, 4, 87–105.

Noussair, C. N. and S. Tucker (2016): "Cash inflows and bubbles in asset markets with constant fundamental values," *Economic Inquiry*, 54, 1596–1606.

Oechssler, J. (2010): "Searching beyond the lamppost: Let's focus on economically relevant questions," *Journal of economic behavior & organization*, 73, 65–67.

Oechssler, J., C. Schmidt, and W. Schnedler (2011): "On the ingredients for bubble formation: informed traders and communication," *Journal of Economic Dynamics and Control*, 35, 1831–1851.

Palan, S. (2013): "A review of bubbles and crashes in experimental asset markets," *Journal of Economic Surveys*, 27, 570–588.

——— (2015): "GIMS - Software for asset market experiments," *Journal of behavioral and experimental finance*, 5, 1–14.

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

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

Powell, O. and N. Shestakova (2016): "Experimental asset markets: A survey of recent developments," *Journal of Behavioral and Experimental Finance*, 12, 14–22.

Razen, M., J. Huber, and M. Kirchler (2017): "Cash inflow and trading horizon in asset markets," *European Economic Review*, 92, 359–384.

Rose, I. N. (2011): "Poker's black Friday," *Gaming Law Review and Economics*, 15, 327–331.

Schilling, L. and H. Uhlig (2019): "Some simple bitcoin economics," *Journal of Monetary Economics*, 106, 16–26.

Shiller, R. (2017): "What is bitcoin really worth? Don't even ask," *The New York Times Dec*, 15.

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

Smith, V. L., G. L. Suchanek, and A. W. Williams (1988): "Bubbles, crashes, and endogenous expectations in experimental spot asset markets," *Econometrica: Journal of the Econometric Society*, 1119–1151.

Smith, V. L., M. Van Boening, and C. P. Wellford (2000): "Dividend timing and behavior in laboratory asset markets," *Economic Theory*, 16, 567–583.

Stöckl, T., J. Huber, and M. Kirchler (2010): "Bubble measures in experimental asset markets," *Experimental Economics*, 13, 284–298.

Thurstone, L. L. (1927): "A law of comparative judgment." *Psychological Review*, 34, 273.

Tucker, S. and Y. Xu (2020): "Nonspeculative Bubbles Revisited: Speculation Does Matter," *Working Paper*.

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

Van Essen, M. and J. Wooders (2015): "Blind stealing: Experience and expertise in a mixed-strategy poker experiment," *Games and Economic Behavior*, 91, 186–206.

Weitzel, U., C. Huber, F. Lindner, J. Huber, J. Rose, and M. Kirchler (2018): "Bubbles and financial professionals," *Innsbruck Working Papers in Economics and Statistics*, 2018-4.

Yellott Jr, J. I. (1977): "The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution," *Journal of Mathematical Psychology*, 15, 109–144.

Yermack, D. (2015): "Is Bitcoin a real currency? An economic appraisal," in *Handbook of digital currency*, Elsevier, 31–43.

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