

Forthcoming. In *Myths and facts about football: The economics and psychology of the world's greatest sport*. P. Andersson, P. Ayton and C. Schmidt. eds. Cambridge: Cambridge Scholars Press.

CHAPTER SEVENTEEN

ACCURACY, CERTAINTY AND SURPRISE - A PREDICTION MARKET ON THE OUTCOME OF THE 2002 FIFA WORLD CUP

**By Carsten Schmidt, Martin Strobel and Henning Oskar
Volkland¹**

The suspicion is that for all the caginess of the odd-setters, this year's Brazil are 1998's Germany—a team in terminal decline (Steve Davies, *Racing Post*, May 28, 2002).

If you can ever write off the Germans, it is this sorry bunch (Steve Davies, *Racing Post*, May 28, 2002).

No hiding place for Guus as Korea attempt to avoid home humiliation (Steve Davies, *Racing Post*, May 28, 2002).

In this chapter, we present our empirical investigation of the forecasting accuracy of a prediction market experiment drawn on the outcome of the World

¹ Schmidt: Sonderforschungsbereich 504, Mannheim University, L 13, 15, 68131 Mannheim, Germany, email: cschmidt@sfb504.uni-mannheim.de. Strobel: Department of Economics, Faculty of Economics and Business Administration, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands, email: m.strobel@algec.unimaas.nl. Volkland: Goldman Sachs & Co., Messeturm, 60308 Frankfurt am Main, Germany, email: oskar.volkland@gs.com. We would like to thank Patric Andersson and Pete Dawson for numerous comments and advice. This chapter is based on a master thesis Volkland has written under the supervision of Strobel. Schmidt revised, reorganised, added additional questionnaire data and shortened the manuscript to provide coherence with the previous chapter and to avoid redundancy. Stata do files of the statistical evaluation and the cdf plots can be made available upon request. Schmidt organised the experiment at the Max-Planck-Institute of Economics, Jena and gratefully acknowledges the financial support from the Max-Planck-Society and the Deutsche Forschungsgemeinschaft (SFB 504). Strobel gratefully acknowledges support from the Dutch Science Foundation (NWO) through the Vernieuwingsimpuls program.

Cup 2002. We analyse the predictive accuracy of 64 markets and compare to bookmakers' quotes and chance as benchmarks. We revisit the evaluation of Schmidt and Werwatz (Chapter 16) and compare our results directly to their findings. In addition, we propose a new method for testing predictive accuracy by means of a non-parametric test for the similarity of probability distributions and we evaluate the incorporation of information in market prices by comparing pre-match and half-time price data.

We find a reversed favourite-longshot bias when analysing market prices before the start of the match and this bias does not disappear with the inflow of new information until half-time. Unlike the market based predictions bookmakers appear to be perfectly calibrated. Since there were substantial deviations in outcome between the 2000 European Championship and our data, we offer possible explanations for the much worse performance of the 2002 World Cup prediction market. Consistent with Schmidt and Werwatz (Chapter 16) prediction markets do assign relatively higher probabilities to the favourite when compared to the odds-setters. Together with a long streak of surprising outcomes this fact appears most likely to be responsible for the predictive inaccuracy.

The chapter empirically tests the efficient market hypothesis. The term was introduced into the economics literature by Hayek. He argued that investors communicate and coordinate their decisions through market prices (Hayek, 1937). The efficient market hypothesis in its most common form claims that financial markets are efficient in the sense that there is no persistent opportunity for abnormal returns based on the observation of past prices or taking into account all public and private information (Fama, 1970). As soon as new information becomes publicly available it is immediately absorbed and incorporated into market prices.

In particular we test the following hypotheses: (1) the markets should predict the outcomes of matches more precisely than chance, (2) the markets should generate more accurate implicit probabilities than the bookmakers' odds and (3) the markets' quotes should be meaningful in a sense that the more certain the market the more often the prediction of the market should be right.

The remainder of this chapter is organised as follows. Section 2 describes the experimental design and data. Section 3 briefly revisits and compares to Schmidt and Werwatz (Chapter 16). Section 4 introduces the new distribution-based test that provides for an alternative test of hypothesis 3. Section 5 looks at explanations for the poor prediction ability of the markets and Section 6 concludes.

The Experiment

The Football Event

The FIFA World Cup 2002 was held in South Korea and Japan. Thirty-two participating teams had qualified for the tournament through a system of regional competitions. The tournament was organised in two stages of 48 (group stage) and 16 (knockout stage) matches, respectively. In the group stage the teams played round robin in eight groups of four to qualify for the knockout stage. The winning team of a match in the group stage received three points; the losing team received zero points. In case of a draw after 90 minutes each team received one point. At the end of the first stage teams were ranked according to the total number of points won from the three group matches. In each group the teams ranked first and second advanced to the knockout stage. In the case that two or more teams obtained the same number of points the direct comparison, i.e. the result of the match against each other, was used as a tie-breaker.² Starting with the knockout stage, a game that was not decided after regular time was continued for a maximum additional time of thirty minutes. The first goal to be scored within this extra time, the so-called ‘golden goal’, decided the game. If a game was still not decided after the additional time, the match outcome would be determined by a penalty shootout. The winner of a game in the knockout stage would progress to the next round.

Experimental Setup

The market experiment used the same rules and software platform as Schmidt and Werwatz (Chapter 16) and was accessible from May 23rd, eight days before the start of the tournament. It remained available until July 2nd, two days after the finals. The first trading day was May 27th, four days before the opening match. Altogether 134 traders participated in the football markets, 72 (54%) were German. The other traders used the English speaking portal. The initial deposits in traders’ accounts ranged from a minimum of 10 Euro to a maximum of 50 Euro.

The number of participants is lower compared to the 2000 European Championship finals but the number of match markets was three times higher. The total commitment in monetary terms was € 3,893 with an average investment of € 29.05. Again, it is a zero sum game—thus no commission is

² Further subordinate tie-breakers are the difference between the numbers of goals scored and received, with the advantage to the team with the higher positive difference, the total number of goals scored in the group stage, the FIFA country coefficient, and, finally, tossing a coin.

charged and all investments will be redistributed and paid back to the participants. Both the total number of trades (19,839) and the average number of trades per participant (148) are five times greater than the corresponding figure for the 2000 European Championship market.

There were two categories of markets that traders could choose to trade in: the championship market and the 64 individual match markets. In the championship market securities issued on each of the 32 participating team were traded. The pay-offs had a winner-takes-all structure: only the contracts of the winning team, the “champion”, paid-off at the end of the tournament, all other contracts expired worthless. The championship market was the only market that remained open for the entire time of the experiment. In the match markets contract design was similar to that of sports bets. During the group stage match market contracts were contingent on one of the three possible outcomes of a match: *first listed team win* (1), *draw* (0) and *second listed team win* (2). In the knockout stage the number of different contracts traded in each market was reduced to two as the contract corresponding to a draw was dropped. The winning contract in a match market, i.e. the contract contingent on the true outcome of a match, yielded a fixed pay-off. All other contracts expired worthless. On the day after a match the liquidation value of their contracts was forwarded to traders through their online accounts. This ensured that traders regained liquidity to reinvest in upcoming markets. The match markets were operated for a limited time prior to a match and closed with the end of this match. Participants could therefore continue to trade in match markets while matches were broadcast live on television. This was not possible in the 2000 European Championship match markets when individual markets were closed at the beginning of their respective game (Schmidt and Werwatz, Chapter 16).

Data

We use professional bookmakers’ fixed odds as a further benchmark for evaluating the predictive accuracy of the market forecasts. For games in the group stage bets can be placed on any of the three outcomes at quotes set by the bookmaker. In the knockout stage the outcome *draw*, or 0, is assigned to games that are not decided after regular time.

We have collected two complete sets of betting odds offered by internet betting agencies. ODDSET data has also been used previously to evaluate the performance of the 2000 European Championship market (Schmidt and Werwatz, Chapter 16). Since ODDSET—which is run by the German state-owned lottery—only allows for German residents to engage in betting, we decided to add the English agency Eurobet to control for potential differences in country specific odds-setting. An obvious difference is due to the differences in

competition: the German quasi-monopolist has a take-out rate of about 25% whereas the English betting agency charges roughly 10%.

We have taken the prices from the match markets prior to the start of the game in order to extract the markets' pre-game predictions. According to the theoretical framework described in the introduction prices in the markets should at all times aggregate all relevant information and expectations about the performance of competing teams. Therefore, the relative prices in a market should reflect opinion about the likelihood that the market assigns to the individual outcomes of a game. Moreover, we collected match markets' prices at half-time. This allows us to test whether new information—based on each team's performance during the first 45 minutes of play—is incorporated in prices and thus implicit probabilities converge to the outcome. In addition, we have collected pre-match prices of the championship market.

The contract with the highest price was selected as the market's predicted winner, i.e. the contract linked to the outcome deemed most likely by the market. We have taken the bets with the lowest quote to be the bookmakers' forecast of a game's outcome. In addition to the qualitative forecasts that the prediction markets and the bookmakers made about the winner of a game, we have also calculated the implicit probabilities assigned to each possible outcome from the market prices and betting odds, respectively.

There were six match markets in the group stage with incomplete trading, in the sense that one or two contracts were not traded at all. For example, in the market for the game Paraguay vs. South Africa there was no trade activity in the contracts *draw* and *second listed team win*, respectively. This was most likely due to a low interest in the match as implied by the low number of trades—8 trades compared to an average number of 38 trades (SD 29)—and the low trade volume—9 Eurocent compared to an overall average trade volume of 39 Eurocent (SD 37)—in the corresponding market. We are aware that assigning not traded contracts the probability 0 might give a flawed view of the market's assessment of the likelihood of the outcomes of the games. Therefore, we also did the empirical analysis by leaving out these matches. Because we did not find significant differences in the results between including and leaving out incomplete matches, the empirical analysis reported in this chapter is based on all 64 matches. This also enables us to directly compare our results with those of Schmidt and Werwatz (Chapter 16).

We observed the same prices for win and lose in the championship market in two matches. To generate a prediction of the outcome we resolve one tie in the championship market (South Korea vs. USA) by assigning the prediction to the contract with the higher pre-match trading volume. Prior to the small final contracts of the contestants South Korea and Turkey were already worthless in the championship market. We use the higher price of both teams' contracts on

midnight prior to the first played of both half-finals to resolve this tie and get Turkey win as prediction.

Besides the three matches with equal odds by ODDSET and the one match with equal odds by Eurobet there has been only one disagreement between the bookmakers (Belgium vs. Russia) on the assignment of the highest win probability (the lowest odd). Therefore, we use in case of equal odds the prediction of the other bookmaker to resolve the tie when predicting the winner. Again, when leaving out matches with equal probabilities from the analysis the results do not change. In addition, we will use in the following section error measures that do not depend on the predicted outcome of the game.

Evaluation of the Predictive (In)Accuracy

Is the Market as Reliable as a Random Predictor?

In the following analysis we will revisit the three hypotheses of Schmidt and Werwatz (Chapter 16) and compare their findings with the results from our data. The first hypothesis makes use of a random predictor, a rolling dice, as a benchmark for the capability of the experimental markets to predict the outcome of uncertain events. We have obtained empirical data on the outcomes of matches in 13 past World Cups from www.fifa.com. The documentation is incomplete for the earlier tournaments in the sense that there is no reference to the first listed team in the official FIFA fixtures. We find 29.5% drawn matches in the group stage and the remaining 70.5% distributed over the two other outcomes. Andersson (Chapter 15) finds the teams listed first to be systematically higher ranked and to win more often compared to the teams listed second. Thus the two outcomes are not equally likely. In order to defend the applicability of the benchmark random dice model, we will use a thought model that randomly assigns the team of record, i.e. the team from which perspectives win or lose is defined. In fact, our following evaluation does not depend on the order of the fixtures. Hence, H_0 states the markets deliver uninformed, random predictions.

Let X_n be the number of correct predictions in n trials. We need to derive the distribution of X_n under the null hypothesis. Since 48 markets were based on three outcomes and 16 markets on two outcomes, $X_n = Y_{48} + Z_{16}$ is the sum of two binomial distributions with $Y_{48} \sim \text{Bin}(48, 1/3)$ and $Z_{16} \sim \text{Bin}(16, 1/2)$. We can reject H_0 in favour of H_1 with a one-sided test at a 5% significance level if we have 31 or more correct predictions.

The match markets correctly predicted the outcome of a match in 34 out of 64 cases. The championship market generated 32 correct predictions. The match markets at half-time (HT) performed best with 35 correct predictions. We can

Table 1. Relative frequency of correct predictions.

	Group stage	Knockout stage ^a	Total
Match market	.50*	.63	.53*
Match market HT	.54**	.56	.54*
Championship market	.46*	.63	.50*
ODDSET	.48*	.50	.48*
Eurobet	.46*	.50	.47*
N	48	16	64

* Significantly different from chance at the 5% level, ** significantly different from chance at the 1% level.

^a Match markets allowed in the knockout stage for betting on *first listed team win* and *second listed team win* only.

reject the null hypothesis that the match market and the championship market is an uninformed, random predictor on a conventional significance level.

Does the Market Beat the Odds?

Compared to the data of the 2000 European Championship markets, the 2002 World Cup markets' and the bookmakers' predictive performance were rather poor. Again, the prediction of the outcome—which is the event having the highest price in the markets and the one possessing the lowest odds with the bookmakers—is not different across markets and bookmakers. ODDSET and Eurobet correctly predicted 31 and 30 out of 64 matches, respectively. It should however be noted that in the knockout stage odd-setters have a disadvantage since they allow for three different outcomes. The distribution of X_n under the null hypothesis now reads $X_{64} \sim Bin(64, 1/3)$ and H_0 can be rejected at a significance level of 5% having 29 or more correct predictions. Thus both odds-setters' predictions are significantly different from the predictions of a random dice. Table 1 provides frequencies of correct predictions for the group stage, the knockout stage and all matches.

Next we make use of the extra information inherent in the magnitude of the markets' predictions in order to evaluate whether the markets were able to generate superior forecasts when compared to the bookmakers as a benchmark. The first measure we employ is the mean squared prediction loss.

$$MSPL_2 = \frac{1}{n} \sum_{i=1}^n [Y_i - P_i^*]^2$$

For each game i the squared prediction error can take on values from 0, if a game is correctly predicted ($Y_i=1$) with probability $P_i^*=1$, to 1, if a game is incorrectly predicted ($Y_i=0$) with probability $P_i^*=1$. The set of predictions that yields the lowest MSPL is hence considered superior. The above formula gives

the definition of the MSPL with 2 outcomes—Schmidt and Werwatz (Chapter 16) provide the definition of the MSPL with three outcomes.

The MSPL is higher in the championship market and the match market when compared to the bookmakers' predictions. The implicit predictions generated from ODDSET's quotes yielded the lowest MSPL of 0.249 with Eurobet relatively close at 0.256. The championship market had the highest error of 0.339 and the match market had an error of 0.300. Only after the first half of games was the match market able to somewhat narrow the gap with the bookmakers; the half-time mean squared prediction loss of the match markets is 0.268.

These results are in contrast to the 2000 European Championship data reported by Schmidt and Werwatz (Chapter 16). In their study the match market was found to be superior to the championship market and to the bookmaker ODDSET in terms of MSPL. The bookmakers' predictive accuracy has been rather constant over the tournaments with a slight improvement for ODDSET from 0.272 in 2000 to 0.249 in 2002.

As a second measure of forecast accuracy we employed the mean logarithmic score (MLS), which has been applied in other studies on prediction markets by, for example, Pennock et al. (2001) and Debnath et al. (2003).

$$MLS = \frac{1}{n} \sum_{i=1}^n \log P_i^R$$

Although both are measures of forecast accuracy, the MSPL and the MLS start from two different focal points. MLS is based on the probability assigned to the ex-post realisation P^R of match i rather than on the ex-ante prediction delivered by the highest probability. Therefore, it does not measure the (in)accuracy of a prediction about the anticipated winners, like MSPL, but rather the (in)accuracy of a prediction about the actual winners.

The MLS is defined with 0 being the maximum and negative infinity the minimum. Higher scores (scores closer to zero) are considered superior. The lower the score's value turns out to be the higher is the surprise of the market or bookmaker about the outcome. The results of our MLS calculation are reported in Table 2. For a comparison we also calculated the corresponding MLS values for the 2000 European Championship: -0.624 for the match markets and -0.980 for ODDSET respectively.³

³ We cannot provide a MLS for the championship markets, since the information obtainable from its prices allows for a statement about the anticipated winner only and not for a complete probability distribution over all three possible outcomes. To be more precise, the relative prices of two teams' contracts in the championship market do yield a probability assigned to the anticipated winner—the team with the higher contract price—but they do not yield a probability assigned to the outcome draw unless the two contracts

Table 2. Mean squared prediction loss (MSPL), mean logarithmic score (MLS) and Brier score (BS).

	MSPL	MLS	BS
Match Market	0.300	-1.068	.629
Match Market half-time	0.268	-0.974	.574
Championship Market	0.339	n.a.	n.a.
ODDSET	0.249	-1.001	.588
Eurobet	0.256	-1.006	.593
N	64	64	64

The market's MLS in 2002 is lower than the MLS for the two bookmakers if we consider all matches. This finding is reversed when compared to the 2000 European Championship data. Again, ODDSET's predictive performance with respect to MLS has been rather constant across tournaments. Thus, both measures, the MSPL and the MLS, consistently provide evidence for predictive inaccuracy of the 2002 World Cup finals prediction market.

In addition, we also computed the Brier score, which is the standard methodology for assessing probabilistic forecasts (see Yates, 1990).

$$BS = \frac{1}{n} \sum_{i=1}^n \sum_{s=1}^S (P_{si} - Y_{si})^2$$

For each possible outcome of the game S (*first listed team win*, *draw* and *second listed team win*) a squared error of its realisation Y_{si} (1 if outcome occurred, 0 otherwise) and its assigned probability P_{si} is aggregated. This parameter is averaged across matches i . Forrest et al. (2005) and Andersson (Chapter 15) report the Brier score for *win*, *draw* and *lose* separately. For a comparison with the Brier scores reported in Table 2 one has to aggregate their Brier score values over the three events.

The magnitude of the brier score is consistent with the results of the MSPL and the MLS. The prediction market before the start of the match performed worst (.629) and improved until half-time (.574). The bookmakers' Brier scores (.588 and .593, respectively) are lower compared to the market before the match. Again, the bookmakers are in a disadvantage by allowing for three outcomes in the 16 matches of the knockout stage compared to the two outcomes of the prediction market.

were traded at the exact same price. There is no such case in the 2002 World Cup market data. Therefore, we leave out this type of markets from the MLS analysis.

Table 3. Logit fit of the determinants of predictive success.

	Match market ^a	Match market HT ^a	Champion- ship market ^a	ODDSET ^a	Eurobet ^a
Constant	-1.167 (1.290)	-2.483 (1.181)*	-2.589 (1.529)	-3.900 (1.477)**	-3.558 (1.344)**
$P_{m,i}^*$	1.887 (1.851)	3.916 (1.704)*	3.161 (1.831)	7.688 (2.914)**	6.678 (2.559)**
N	64	64	64	64	64
Psydo R^2	0.01	0.07	0.04	0.09	0.09

* Significant at the 5% level, ** significant at the 1% level. Standard errors in parentheses.

^a Dependent variable: $Y_{m,i}$: 1—if contract with highest implicit probability predicted the outcome correctly, 0—otherwise.

Determinants of Predictive (In)Accuracy

The next stage of the analysis is to identify variables that influence the accuracy of the market predictions. Assuming some forecasting power in the market predictions we should find that the more certain the market was the more often its predictions were correct. To this end, we follow Schmidt and Werwatz (Chapter 16) in using a logistic regression to quantify the effect of certainty on forecast accuracy for the championship market and the bookmakers. The results are reported in Table 3. It turns out that certainty—measured as the magnitude P_i^* of the contract with the highest probability—has a significant effect on the two bookmakers' predictions. This is not the case with pre-match predictions of the market; only at half-time the degree of certainty has a significant effect. Once again, these findings from the 2002 World Cup prediction market are not in line with the results of the 2000 European Championship market.

A Distribution-Based Test for Predictive (In)Accuracy

In this section we use a novel approach for testing the accuracy of market-generated predictions which tries to integrate the prediction of the outcome and the certainty of the prediction. The test can be seen as an extension of Pennock et al. (2001) who sorted market predictions according to their prediction probability into different buckets. With each of the buckets they ran a binomial test with the average prediction probability as base probability.

Instead of grouping into buckets we order all different contracts—win, draw and lose—according to their assigned probability and give the value 1 to the winning contracts and 0 otherwise. The collection of all different contracts with

their assigned values of 0 or 1 reflects a pseudo probability density function (pdf) of the distribution of the paying-off contracts. If we move along the probability spectrum from 0 to 1 and add up the assigned binary values of the contracts we get a trace of the cumulative frequency of the 64 contracts that actually paid-off. In other words, the cumulative frequency of success mimics the cumulative distribution function (cdf) of the distribution of the 64 paying-off contracts among the total number of different contracts. This pseudo cdf is a step function that makes 64 jumps of constant height of $1/64$, each at a point within the probability spectrum that corresponds to one of the 64 paying-off contracts. In what follows, we will refer to this pdf and cdf as the actual distribution of success, since it mirrors the way the tournament has evolved ex-post. If predictions were accurate, then from the segment of the probability spectrum around .10 about one in ten contracts should have won, two out of ten contracts around .20, and so on. We call this the actual cdf.

A second distribution will be called the theoretical distribution, since it mirrors the way that the prediction markets and the bookmakers have theoretically (ex-ante) anticipated the tournament to evolve. We assign to each different contract the value of its market- or odds-setter determined probability. Naturally, the resulting cdf has many more jumps than the actual cdf, each of unequal height. A jump occurs at each value from the probability spectrum of the three (two) different outcomes of the 64 events. The height of this jump is determined by the value of the probability assigned to the particular contract. Predictive accuracy is positively related to similarity in shape and, hence, the closeness of the actual and the theoretical cdf.

We introduce a third distribution in order to reconsider some results from earlier sections that used the random predictor as a benchmark for predictive accuracy. It assumes that there is absolutely no predictive power in markets' and bookmakers' implicit probabilities. We call it the agnostic distribution, since it assigns the same probability to all different contracts. The height of the jumps of the agnostic cdf is constant at the value $1/N$, with N being the total number of different contracts. Since the markets allow for two different contracts in the knockout stage and the bookmakers for three we get $N=176$ for the markets' cdf and $N=192$ for the bookmakers' cdf. Again, each jump occurs at the market or bookmaker assigned probability.

We observe that the distribution of the probabilities the markets assigned to events ranges from 0 to 1—when not considering the 6 matches with incomplete contracts the range is 0.01 to 0.97—whereas the distribution of the bookmakers' implicit probabilities is less extreme. Eurobet issued quotes with an implicit probability from 0.04 to 0.82 whilst ODDSET seems to publish the most conservative quotes with values ranging from 0.08 to 0.73 (compare x-axis, Figure 1).

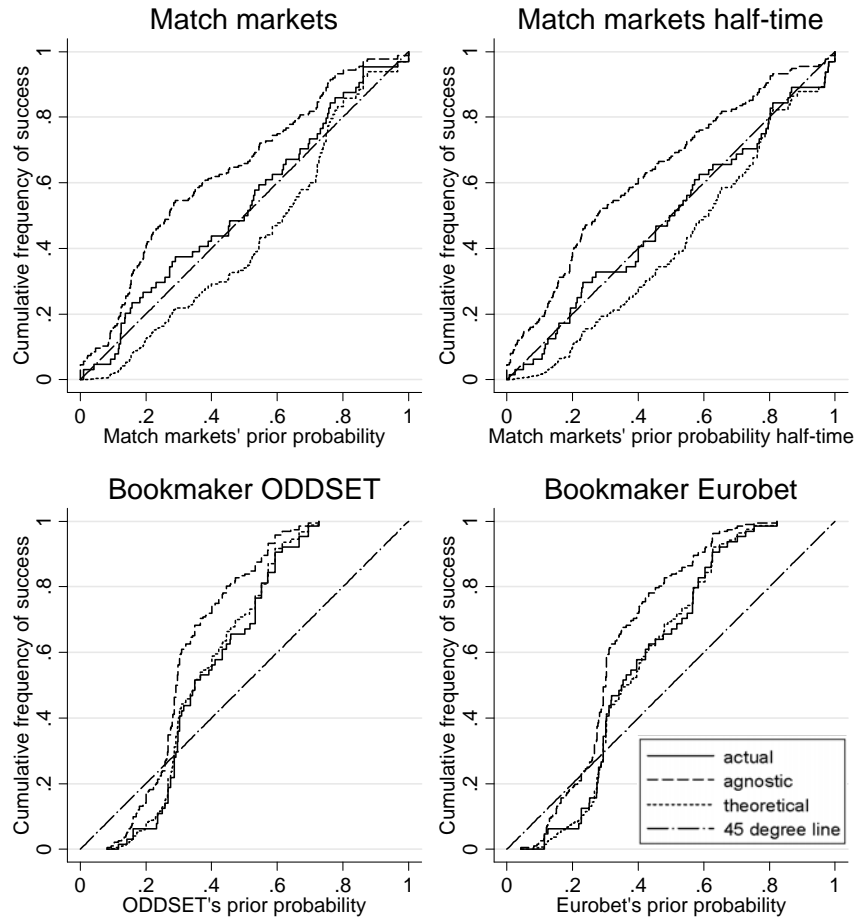


Figure 1. Actual-, theoretical- and agnostic cdf of the distribution of successes. Predictive accuracy is positively related to similarity in shape and closeness of the theoretical and the actual cdf.

The convex shape of the markets' theoretical cdf displayed in Figure 1 in both upper panels provides evidence for a too high concentration of implicit probabilities at the very high and very low end of the probability spectrum. Furthermore, this high degree of certainty in the markets' assigned probabilities implies a relatively flat increase of the cdf over the middle range of probability. This appears to be because extreme positions tend to drive out conservative

Table 4. Kolmogorov-Smirnov test of the similarity of the cumulative distribution.

Hypothesis	Match market ^a	Match market HT ^a	ODDSET ^a	Eurobet ^a
1 theoretical vs. agnostic	0.341***	0.341***	0.177**	0.193***
2 actual vs. agnostic	0.210***	0.256***	0.224***	0.198***
3 theoretical vs. actual	0.296***	0.205***	0.068	0.089
N	176	176	192	192

** Significant at the 1% level, *** significant at the 0.1% level.

^a Kolmogorov-Smirnov test statistic D

mid-range bets. As a consequence, the degree of certainty in the markets influences the degree of convexity of the theoretical cdf.

In other words the market participants systematically over-estimated very likely outcomes and under-estimated very unlikely outcomes. This observation provides evidence for the presence of a reversed favourite-longshot bias in markets' prices, stating that bettors overestimate the favourites' probability of winning and underestimate the longshots' probability of winning. The reversed favourite-longshot bias can be observed for the markets at half-time as well meaning that the inflow of new information does not drive out the bias. This bias cannot be observed for the bookmakers' quotes.

In the next step we formally test the closeness and similarity of shape of the three different distributions. If the realisation of the football tournament is not different from a rolling dice we should not find a significant difference between the agnostic and the actual cumulative distribution. If the markets and bookmakers predict well, we should not find significant differences between the theoretical and the actual cumulative distribution. If the markets and bookmakers do not take any information into account we should not find significant differences between the theoretical and the agnostic distribution. We formulate the following hypotheses:

H1: The market's/bookmakers' prediction is not systematically different than would be predicted by a random predictor (theoretical vs. agnostic).

H2: The actual realisation is not systematically different than would be predicted by a random predictor (actual vs. agnostic).

H3: The market's/bookmakers' prediction is not systematically different from the actual realisation (theoretical vs. actual).

We use a Kolmogorov-Smirnov equality-of-distributions test.⁴ The results of the hypothesis tests are provided in Table 4. We can reject the first hypothesis

⁴ We are aware of the fact that the three (two) contracts of a match are not iid and significance levels might be inflated. An alternative way to proceed is to consider one of

for all sources and find that the market prediction and the bookmakers' quotes do differ from a random predictor. We can reject the second hypothesis and find that the actual realisation of the tournament's outcomes is different from rolling a dice. Finally, we are unable to reject the third hypothesis for the two bookmakers, thus we are not able to find significant differences between bookmakers' predictions and actual realisation. This is not the case with the markets, which observe significant differences between the distribution of the prediction and actual realization. The two lower panels of Figure 1 visualise the test results: bookmakers' actual and theoretical cdf appear to be quite similar across the full range of the probability spectrum. The bookmakers seem to be perfectly calibrated, whereas the markets are biased when considering very low and very large probabilities. In general, the rather close correspondence between the test results from our new method with the findings of the previous sections mutually confirms our results.

Two Explanations for the Predictive (In)Accuracy

Prior beliefs

The rather poor predictive accuracy appears to be evidence against the functioning of the markets as an information aggregation mechanism that is able to accurately predict the outcomes of uncertain events. However, previous studies on laboratory experiments and prediction markets have, in the majority of cases, presented highly convincing results in favour of the information aggregation potential of the markets (Plott and Sunder, 1988, Sunder, 1995, Hanson, 1998; Gruca et al., 2003). Besides, a substantial fraction of traders in the 2002 markets had already participated in the 2000 European Championship experiment. This casts further doubt on the idea of a structural problem of the markets' design.

A remaining hypothesis is that the 2002 World Cup markets were able to accumulate and to process the information held by the participants, but that traders were just collectively wrong too often. Gruca et al. (2003) have used a prediction market in a laboratory experiment in order to examine the relationship between the aggregation of trader information and the accuracy of a

the three outcomes only. Since the FIFA fixtures of *first listed team win* and *second listed team win* are not random it is not possible to provide a unique agnostic cdf. We have tested hypothesis 3 for all three possible outcomes separately. We find the D statistic of the K-S test in the match market (at half-time) for *first listed team win* 0.297** (0.156), *draw* 0.271* (0.229) and *second listed team win* 0.219⁺ (0.266*). There are no significant differences when comparing the distribution of the cdf of actual outcome to bookmakers' predictions.

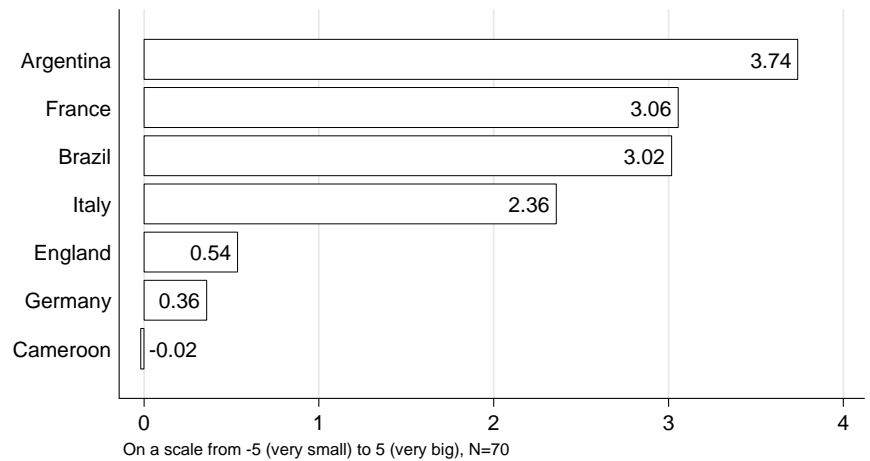


Figure 2. “What is the likelihood a team will win the championships?” Selected teams from the self-reported subject questionnaire filled out at the first login and prior to the start of the tournament.

forecast of new product success. The authors find a high degree of effectiveness in aggregating information but a relatively bad forecast accuracy. They could not verify a strong relationship between the degree of aggregation and forecasting accuracy and therefore promote a careful distinction between these two outwardly similar concepts.

In order to determine whether traders were collectively wrong we will evaluate the a priori expectations of traders. A questionnaire was presented to the subjects when they logged in for the very first time. Since we decided to ask for prior beliefs before the first match only, we are left with responses from 70 out of 134 traders. We asked them to rate the probability a team will be the overall champion from -5 (very small) to 5 (very big). Figure 2 displays selected teams.

It turns out that not many of the traders’ prior beliefs were confirmed by subsequent match outcomes. The traders in our experiment appear to have had difficulties in judging the prospects of seemingly strong teams. The most prominent example of a “surprise team” was the defending champion France, who failed to score in their three group matches and who consequently did not qualify for the knockout stage. The probability of a victory for France in the opening match against Senegal was set at 75% by the match market before kick-off. For the record, France took its assigned 9% chance and lost this game. One might ask whether the traders reacted upon the observed performance of a team over the course of the tournament and adjusted their expectations accordingly.

Table 5. Odds for the 2002 World Cup winner published three days in advance of the start of the tournament.

	Bet365	Bet Direct	Coral	Hills	Ladbrokes
France	7-2	3	4	3	3
Argentina	5	9-2	9-2	9-2	9-2
Italy	9-2	5	5	5	5
Brazil	11-2	6	6	11-2	6
England	14	14	10	9	9
Germany	14	12	14	16	20
Cameroon	33	40	40	25	33

Source: *Racing Post* World Cup 2002 Preview Pull-out. May 28, 2002.

However, a second look at the probability assigned to a victory of the obviously weak French in the two subsequent group matches does not imply such expectation-revising behaviour: it was 77% in the match against Uruguay (final outcome 0:0, 14%), and 73% in the match against Denmark (final outcome 0:2, 12%). The market remained uncritically certain that the French would still qualify from the group.

There is, however, another aspect that supports the idea of a collectively wrong perception about the matches and their likely outcomes that does not contradict our assumption of a well functioning information aggregation mechanism. In more than 90% of the games the predictions of the match markets as well as bookmakers hinted at the exact same outcome. Recall that the bookmakers in our sample represent experts from two different countries. Since the numbers of correctly predicted outcomes are also more or less stable across prediction markets and bookmakers we might say that traders and experts did all share essentially the same a priori information and both equally surprised by the final outcomes.

On May 28th, 2002, three days before the start of the tournament, the British sports betting journal *Racing Post* published a pull-out with all tournament related bets offered by British betting agencies. Total betting turnover was said to dwarf any other sporting event with an estimated £200m staked in the UK, twice the 2000 European Championship total.⁵ Given this enormous financial outlay we believe that UK betting agencies qualified as experts on the upcoming event. Table 5 shows a selection of the quotes of the tournament's extended set of favourites published in the pull-out in order to sketch the a priori consensus opinion of most experts. There is no disagreement between bookmakers and only slight disagreement between market participants and bookmakers. For example market traders identified Argentina as the tournament favourite and

⁵ In *Racing Post* World Cup 2002 Preview Pull-out. May 28, 2002.

Table 6. Correlations of the degree of certainty measured as probability of the favourite across market's and bookmakers' predictions.

Probability of win by the favourite P_i^* Correlations (lower part)	Match market	Match market half-time	Championship market	ODDSET	Eurobet
Mean (SD)	.67 (.13)	.67 (.16)	.82 (.15)	.50 (.10)	.52 (.12)
Median	.71	.67	.86	.51	.49
N	64	64	64	64	64
Match market	1				
Match market half-time	.67	1			
Championship market	.50	.32	1		
ODDSET	.53	.33	.66	1	
Eurobet	.43	.25	.62	.95	1

bookmakers the then-defending champion France. Nevertheless, neither France nor Argentina did well in the 2002 World Cup. To this end we are convinced that there was a big surprise element about the actual outcomes among bookmakers and traders.

Biased quantitative predictions, preferences for the favourite and commissions

The next part of the investigation is to establish the markets' relatively poor accuracy with respect to the bookmakers' quotes on a quantitative level. The bookmakers do not appear to have been able to aggregate more or "better" information given their equally low numbers of correctly predicted outcomes. The markets and the bookmakers predicted the same outcomes in more than 90% of the matches. As a consequence, their respective number of incorrect predictions is also approximately the same. Nevertheless, the markets' MSPL is considerably higher than bookmakers' prediction error. Recall that the MSPL rewards and punishes a high degree of certainty in case of a correct prediction and in case of a false prediction, respectively. With the many incorrect predictions at hand, a more conservative approach is likely to have been rewarded. Therefore, we believe that the prediction markets on football are more certain by assigning higher probabilities to their anticipated winners. This effect should be strongest in the championship market that has produced the highest MSPL over the whole tournament (see Table 3). However, it should be intuitively clear that the average certainty is the highest in the championship market, since its win-lose prediction guarantees a minimum probability of .5 to an anticipated winner. The bookmakers on the other hand have been less certain

Table 7. Number of games and MSPL forecast error by correctly and incorrectly predicted games, European Championship 2000 experiment for comparison.

	Prediction	World Cup 2002		European Championship 2000 ^a	
		N	MSPL	N	MSPL
Match market	Correct (Incorrect)	34 (30)	.07 (.55)	15 (6)	.08 (.39)
Match market HT	Correct (Incorrect)	35 (29)	.06 (.51)	-	-
Championship market	Correct (Incorrect)	32 (32)	.04 (.64)	14 (7)	.08 (.43)
ODDSET	Correct (Incorrect)	31 (33)	.13 (.36)	13 (8)	.18 (.37)
Eurobet	Correct (Incorrect)	30 (34)	.12 (.38)	-	-

^aSource: Schmidt and Werwatz (Chapter 16).

about their favourites and avoided assigning extremely high (or low) probabilities to any single outcome.

Table 6 provides some insight into the differences in predictions and certainty from a comparison of the favourites' probabilities across prediction markets and bookmakers. The two bookmakers exhibit roughly the same degree of certainty and a high correlation. The pre-match and half-time predictions from the match market are also quite close to each other, although their correlation is less than that of the bookmakers. As expected, the certainty about its predictions was highest in the championship market.

The implication of the high degree of certainty in the markets can be observed from the comparison of the prediction error produced in correctly and incorrectly predicted games (see Table 7). Whenever the markets' predictions were correct they yielded an extremely low forecast error (0.07, 0.06 and 0.04 per correct game), much lower than ODDSET and Eurobet (0.13 and 0.12 per correct game). Whenever the predictions were incorrect, however, the opposite is true: the markets produced a per-game forecast error that was much higher than the corresponding odds-setter's error (0.55, 0.51 and 0.64 vs. 0.36 and 0.38). This pattern is roughly the same as in the 2000 European Championship experiment. The high degree of certainty that was the greatest strength of the markets in the less surprising 2000 situation has become their greatest weakness in the highly surprising 2002 tournament.

We now consider a possible explanation for the weak calibration of the predictions' generated from the 2002 World Cup market. The literature has attributed biases in predictions to be due to a preference for the favourite and differences in commission (Williams and Panton, 1998). The so called reversed favourite-longshot bias has been observed in the baseball fixed odds betting market (Woodland and Woodland, 1994, 2003), which is characterised by comparatively low commissions. Smith et al. (2006) find a reduced favourite

longshot bias in exchange based betting when they compared to fixed odds in horse racing. Preferences for favourites on the demand side have been reported by Levitt (2004) who finds bettors to have a higher demand for wagers on the favourites in American football fixed-spread betting.

The high degree of certainty by market participants with respect to their favourites goes hand in hand with a low degree of certainty for longshots (see Figure 1). Together with the differences in commissions—the markets did not charge any commission whereas ODDSET's and Eurobet's take is substantial—the observed reversed favourite-longshot bias with the markets is in line with the predictions of Williams and Panton (1998).

Conclusions

In this chapter we have evaluated the predictive performance of experimental asset markets that were conducted during the 2002 World Cup. Contracts in the markets were contingent on the outcomes of football matches. According to the theory of efficient markets and rational expectations, contract prices from the markets should be the best available forecast for the outcome of a match since they reflect the consensus opinion of all traders about the match. Previous work in the field, in particular the study of Schmidt and Werwatz (Chapter 16) based on the 2000 European Championship, found a predictive performance of markets that was superior to forecasts from odds-setters.

In this chapter we find market performance and bookmaker performance to be significantly different from a random predictor. In contrast to the findings from the preceding 2000 European Championship experiment, the market's predictions were less accurate than the predictions from two expert bookmakers. In addition, the certainty of the market's prediction was not significant in explaining the predictive success of the markets. In fact, no explanatory factor could be identified that was significant in explaining predictive accuracy by the markets. We argue that the nature of the underlying sports event is a key element in explaining the deviations in outcome between the two studies. The aggregation of all relevant a priori information about a sports event may well be a feasible task for markets of our type, but that does not imply that the resulting forecast is necessarily of superior accuracy. The 'human' factor remains too high in the football game and does at times lead to surprising outcomes as was quite apparent in the 2002 World Cup. Consistently with Schmidt and Werwatz (Chapter 16) we found the implicit probabilities derived from market prices to be more certain about the outcome than the probabilities derived from bookmakers' odds. This certainty was met by the many surprising outcomes and has resulted in the poor predictive performance by the markets. The bookmakers on the other hand took much less extreme stakes and were not hit as hard by the

surprising outcomes. Betting behaviour and the degree of certainty of the implicit probabilities generated from market prices and bookmakers' odds were remarkably alike across experiments, and we therefore favour surprise as the chief explanation for the difference in findings.

The favourite-longshot bias, stating that bettors overestimate the longshots' probability of winning, is well recorded in the racetrack literature. It has recently been investigated in another Iowa Electronic Market type experiment (Berg and Rietz, 2002). The authors find in their study that there exists a reverse favourite-longshot bias that is attributable to overconfidence among traders. As a consequence, contracts with relatively bad win prospects are priced significantly too low and vice versa. The price data from the 2002 World Cup markets is consistent with this finding.

We further developed and implemented a new approach to test for the accuracy of market-generated predictions. Specifically we used a Kolmogorov-Smirnov test for a comparison of the empirical distribution function of match outcomes with the distribution of all prediction market and bookmaker predictions. It turned out that the distribution of the market predictions was significantly different from the distribution of the tournament's actual outcome. This is not the case with both bookmakers who are perfectly calibrated in the 2002 World Cup. Nonetheless, the market predictions (and the bookmakers' predictions) were significantly different from a random predictor. We consider these results to provide further support for our previous explanations.

Further research should aim for more participation in the markets that will generate more data for more sophisticated analyses. In particular, the new opportunity to keep trading during matches has generated limited additional data in the 2002 World Cup experiment. Thus, we could not have made any meaningful analysis of how prices react to events during a game. Gil and Levitt (2008) use intrade data from the 2002 World Cup and study contracts while the game is played. They find mixed evidence for market efficiency in this market: prices react to goals, yet prices continue to trend higher for 10 to 15 minutes after the goal. We were able to show that in the context of the 2002 World Cup prediction market prices react with respect to predictive accuracy to new information inflow until half-time. Yet, this new information at half-time is not able to drive out the reverse favourite-longshot bias observed with market prices in the 2002 World Cup prediction market.

Predicting the outcome of football matches is a highly uncertain task. An empirical argument was put forward by Wagenaar (1988), who argued that transitivity with respect to the match outcome among teams in World Cup games is rare. For the 2002 World Cup competition Group D might serve as an illustrative example where the transitivity of outcomes is violated: USA won over Portugal, Portugal over Poland, yet Poland succeeded over the USA. One

underlying reason why chance plays such a big role in football outcomes might be the game's low scoring property. It remains an open issue for further empirical investigations whether the degree of surprise in the 2002 World Cup is the norm for World Cup and European Championship matches or an outlier.

References

- Andersson, P. (2008). "Expert Predictions of Football: A Survey of the Literature and an Empirical Inquiry into Tipsters' and Odds-Setters' Ability to Predict the World Cup." In *Myths and facts about football: The economics and psychology of the world's greatest sport*. P. Andersson, P. Ayton and C. Schmidt. eds. Cambridge: Cambridge Scholars Press.
- Berg, J. and T. Rietz (2002). *Longshots, Overconfidence and Efficiency on the Iowa Electronic Market*. Working paper, University of Iowa.
- Debnath, S., D. Pennock, S. Lawrence, E. Glover, and C. Giles (2003). "Characterizing Efficiency and Information Incorporation in Sports Betting Markets." In *Proceedings of the Fourth Annual ACM Conference on Electronic Commerce (EC'03)*.
- Fama, E. F. (1970). "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance*, 25(2):383-417.
- Forrest, D., J. Goddard and R. Simmons (2005). "Odds-Setters as Forecasters: The Case of English Football." *International Journal of Forecasting*, 21(3), 551-564.
- Gil, R. and S. D. Levitt (2008). "Testing the Efficiency of Markets in the 2002 World Cup." *The Journal of Prediction Markets*, 1 (3): 255-270.
- Gruca, T., J. Berg, and M. Cipriano (2003). *Limits to Information Aggregation in Electronic Prediction Markets*. Working paper, University of Iowa.
- Hanson, R. (1998) "Consensus by Identifying Extremists." *Theory and Decision*, 44(3):293-301.
- Hayek, F. (1937). "Economics and Knowledge." *Economica*, NS 4, 33-54, 1937.
- Levitt, S. D. (2004). "Why are Gambling Markets Organised so Differently from Financial Markets?" *The Economic Journal*, 114: 223-246.
- Pennock, D., S. Lawrence, C. Giles, and F. Nielsen (2001). "Extracting Collective Probabilistic Forecasts from Web Games." In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*.
- Plott, C. and S. Sunder. (1998). "Rational Expectations and the Aggregation of Diverse Information in Laboratory Security Markets." *Econometrica*, 56(5):1085-1118.
- Schmidt, C. and A. Werwatz. (2008). "How Accurately Do Markets Predict the Outcome of an Event? The Euro 2000 Football Experiment." In *Myths and*

- facts about football: The economics and psychology of the world's greatest sport.* P. Andersson, P. Ayton and C. Schmidt. eds. Cambridge: Cambridge Scholars Press.
- Smith, M. A., W. D. Paton and L. V. Williams (2006). "Market Efficiency in Person-to-Person Betting." *Economica* 73, 673–689.
- Sunder, S. (1995). "Experimental Asset Markets: A Survey." In J. Kagel and A. Roth, editors, *The Handbook of Experimental Economics*, Princeton University Press, 415-500.
- Wagenaar, W. A. (1988). *Paradoxes of Gambling Behavior*. Hove: Lawrence Erlbaum Associates.
- Woodland, L. M and Woodland, B. M. (1994). "Market Efficiency and the Favourite-Longshot Bias: The Baseball Betting Market." *Journal of Finance*, 49: 269-279.
- Woodland, L. M and Woodland, B. M. (2003). "The Reverse Favourite-Longshot Bias and Market Efficiency in Major League Baseball: An Update." *Bulletin of Economic Research*, 55(2): 113-123.
- Williams, L. V. and D. Panton (1998). "Why are Some Favourite-Longshot Biases Positive and Others Negative?" *Applied Economics*, 30: 1505-1510.
- Yates, J .F. (1990). *Judgment and Decision Making*. Englewood Cliffs, NJ: Prentice Hall.