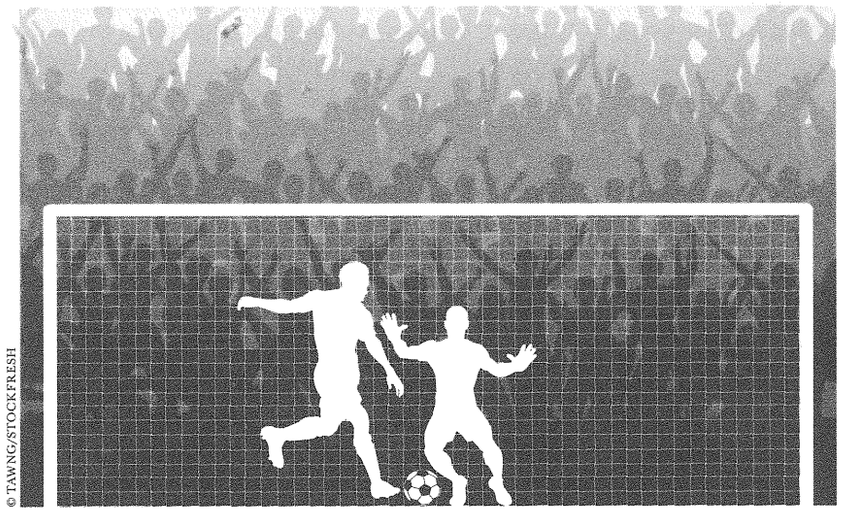
1

PELÉ MEETS JOHN VON NEUMANN IN THE PENALTY AREA



© TAWNG/STOCKFRESH

I thought there was nothing worth publishing until the Minimax Theorem was proved. As far as I can see, there could be no theory of games without that theorem.
—John von Neumann 1953

Much real-world strategic interaction cannot be fully understood with current tools. To make further progress, the field needs to gain more experience in applications to the real world.
—Game Theory Society 2006

The Hungarian national soccer team of the 1950s was one of the greatest soccer teams in the history of the 20th century. It played against England at Empire Wembley Stadium on November 25, 1953, in front of 105,000 people, in what was termed "the match of the century." Hungary was the world's number one ranked team and on a run of 24 unbeaten games. England was the world's number three ranked team, and unbeaten at Wembley for 90 years against teams from outside the

British Isles. In what was then considered a shocking result, Hungary beat England 6–3.

As a preparation for the 1954 World Cup in Switzerland, on May 23, 1954, England visited Budapest in the hope of avenging the 6–3 defeat. Instead, Hungary gave another master class, beating England 7–1. This score still ranks as England's worst defeat.

In those years, soccer was already the world's most popular sport, and Hungary was the best soccer team in the world—often considered one of the four or five best teams in history. As Olympic champion in 1952, not surprisingly, Hungary was the favorite to win the upcoming World Cup. Consistent with these expectations, Hungary easily beat Korea 9–0 and West Germany 8–3 in the first round. Then, it beat Brazil 4–2 in the quarterfinals and Uruguay, which had never been beaten in World Cup games, 4–2 in the semifinals. Its opponent in the final was West Germany, which surprisingly had managed to win all of its games after its initial defeat in the first round to Hungary. In the Wankdorf Stadium in Bern, 60,000 people saw Germany beat Hungary 3–2 in what was called "the Miracle of Bern." The sports announcer shouted in the background of the final scene of Rainer Werner Fassbinder's film *The Marriage of Maria Braun,* featuring this event, "Deutschland ist wieder was!" (Germany is something again!). This victory represented a powerful symbol of Germany's recovery from the ravages of the Second World War.

It is probably safe to assume that there were few Hungarians in the world in the 1950s who were not proudly aware of the accomplishments of their national team in the world's most popular sport. Indeed, Neumann Janos, born on December 28, 1903, in Budapest, a superb scientist who had migrated to the United States and was then known as John von Neumann, could not have been completely ignorant of the team's success.

John von Neumann is considered a scientific genius of the 20th century. A brilliant mathematician and physicist, he left a profound mark, with fundamental contributions in theoretical physics, applied physics, decision theory, meteorology, biology, economics, and nuclear deterrence; he became, more than any other individual, the creator of the modern digital computer.

"He was a genius, the fastest mind I have ever encountered. . . . He darted briefly into our domain, and it has never been the same since." Paul Samuelson (Nobel laureate in Economics, 1970, quoted in Macrae 1992) is referring here to the three fundamental contributions von Neumann made in economics: first, his 1928 paper "Zur Theorie der Gesellschaftsspiele," published in *Mathematische Annalen*, which established von Neumann as the father of game theory; second, his 1937 paper "A Model of General Equilibrium" (translated and published in 1945–46 in the *Review of Economic Studies*); and third, his classic book

Theory of Games and Economic Behavior, coauthored with Oskar Morgenstern in 1944 (Macrae 1992).

As a mathematician, von Neumann's own philosophical views induced him to choose to work in a variety of fields, and his selection of research questions and the resulting measure of his success were largely influenced by aesthetic values (von Neumann 1947). However, he also warned that mathematics loses much of its creative drive when too far removed from empirical sources. And yet, despite the place in the world of soccer that Hungary occupied, and despite soccer's place as the world's most popular game, everything indicates that he was not particularly interested in sports as an empirical source, or in the empirical verification of game theory theorems with sports data or with any data:

> The truth is that, to the best of my knowledge, my father had absolutely no interest in soccer or any other team sport, even as a spectator or news-follower. Ironically, he wasn't much on games in general (though he loved children's toys, which he could often persuade to yield up some principle of mathematics or physics); but his game-playing didn't go much beyond an occasional game of Chinese checkers at my request. I don't believe he even played poker.
>
> Warmest regards, Marina von Neumann Whitman
> (Private email correspondence, October 13, 2010)

As Kreps (1991) notes, "the point of game theory is to help economists understand and predict what will happen in economic, social and political contexts."[1] But if von Neumann considered, as the initial quotation suggests, that there could be no theory of games without proving the minimax theorem, then it seems appropriate to think that he would have considered that there could be no empirical applicability of the theory of games without first having verified empirically that theorem. As noted below, the minimax theorem was not empirically verified until 2003.

The empirical verification of strategic models of behavior is often difficult and problematic. In fact, testing the implications of *any* game theoretical model in a real-life setting has proven extremely difficult in the economics literature for a number of reasons. The primary reason is that many predictions often hinge on properties of the utility functions and the values of the rewards used. Furthermore, even when predictions are invariant over classes of preferences, data on rewards are seldom available in natural settings. Moreover, there is often great difficulty in determining the actual strategies available to the individuals involved, as well as in measuring these individuals' choices, effort levels, and the

1 Not everyone agrees that this is the point. See Rubinstein (2012).

incentive structures they face. As a result, even the most fundamental predictions of game-theoretical models have not yet been supported empirically in real situations.

Von Neumann published the minimax theorem in his 1928 article "Zur Theorie der Gesellschaftsspiele." The theorem essentially says,

> For every two-person, zero-sum game with finitely many strategies, there exists a value V and a mixed strategy for each player, such that:
>
> (a) Given player 2's strategy, the best payoff possible for player 1 is V, and
> (b) Given player 1's strategy, the best payoff possible for player 2 is $-V$.
>
> Equivalently, Player 1's strategy guarantees him a payoff of V regardless of Player 2's strategy, and similarly Player 2 can guarantee himself a payoff of $-V$.

A mixed strategy is a strategy consisting of possible moves and a probability distribution (collection of weights) that corresponds to how frequently each move is to be played. Interestingly, there are a number of interpretations of mixed strategy equilibrium, and economists often disagree as to which one is the most appropriate. See, for example, the interesting discussion in the classic graduate textbook by Martin Osborne and Ariel Rubinstein, *A Course on Game Theory* (1994, Section 3.2).

A game is called zero-sum or, more generally, constant-sum, if the two players' payoffs always sum to a constant, the idea being that the payoff of one player is always exactly the negative of that of the other player. The name "minimax" arises because each player minimizes the maximum payoff possible for the other. Since the game is zero-sum, he or she also minimizes his or her own maximum loss (i.e., maximizes his or her minimum payoff).

Most games or strategic situations in reality involve a mixture of conflict and common interest. Sometimes everyone wins, such as when players engage in voluntary trade for mutual benefit. In other situations, everyone can lose, as the well-known prisoner's dilemma situations illustrate. Thus, the case of *pure conflict* (or zero-sum or constant-sum or strictly competitive) games represents the extreme case of conflict situations that involve no common interest. As such, and as Aumann (1987) puts it, zero-sum games are a "vital cornerstone of game theory." It is not a surprise that they were the first to be studied theoretically.

The minimax theorem can be regarded as a special case of the more general theory of Nash (1950, 1951). It applies only to two-person, zero-sum or constant-sum games, whereas the Nash equilibrium concept can be used with any number of players and any mixture of conflict and common interest in the game.

Before undertaking a formal analysis, let us take a brief detour and look at the following play in soccer: a penalty kick. A penalty kick is awarded against a team that commits one of the 10 punishable offenses inside its own penalty area while the ball is in play. The world governing body of soccer, the Fédération Internationale de Football Association (FIFA), describes the simple rules that govern this play in the official *Laws of the Game* (FIFA 2012). First, the position of the ball and the players are determined as follows:

- "The ball is placed on the penalty mark in the penalty area.
- The player taking the penalty kick is properly identified.
- The defending goalkeeper remains on the goal line, facing the kicker, between the goalposts, until the ball has been kicked.
- The players other than the kicker are located inside the field of play, outside the penalty area, behind the penalty mark, and at least 10 yards (9.15 meters) from the penalty mark."

Then,

- "The player taking the penalty kicks the ball forward.
- He does not play the ball a second time until it has touched another player.
- A goal may be scored directly from a penalty kick."

The credit of inventing this play belongs to William McCrum. McCrum was a wealthy linen manufacturer, raconteur, cricketer, and the goalkeeper of Milford Everton, a small club in County Armagh, which played the inaugural season of the Irish Championship in 1890–91. History does not fully record how good a keeper he was, but he was certainly kept busy during that first Irish League season. Milford Everton finished at the bottom of the league with no points, a record of 10 goals scored, and 62 conceded, and the team was promptly relegated. McCrum may not have been one of the world's greatest goalkeepers, but he was a gentleman and justly proud of his reputation for good sportsmanship. His obituary in 1932 paints a picture of a man of honor who was frustrated and angry at the "win-at-all-costs" attitude that was poisoning his beloved soccer (Miller 1998).

McCrum believed that anyone who failed to abide by the spirit of the game should face a sanction that would punish not just the individual offender but also the whole team. Holding an influential position in the Irish Football Association, he submitted his proposal for a "penalty kick" to the association in 1890. Jack Reid, general secretary of the association, then formally forwarded McCrum's proposal to the international board for consideration at the board meeting to be held on June 2, 1890, and, he hoped, its subsequent incorporation into the laws. It immediately ran into a storm of protest. The reception was ferocious.

Press, administrators, and players publicly derided the idea. Some commentators even nicknamed the proposal the "death penalty," implying that it would be the death of the game as they knew it. Many people did not want to introduce a rule that effectively conceded that teams and players often resorted to unsporting methods. It was in this atmosphere that the Irish Football Association decided to withdraw the proposal. The international board, however, agreed to discuss the issue at the next meeting one year later. On June 2, 1891, somewhat surprisingly, the atmosphere was quite different and the proposal passed unanimously.

The penalty kick was born, albeit not in the form that we know it today. The new law came into force immediately, and, to be fair, it was not a huge success. There were obvious flaws in the first draft, and players—particularly goalkeepers—were quick to take advantage. Furthermore, gentlemen did not commit fouls. It took almost 40 years, until 1929, before the penalty law finally became what William McCrum intended it to be—an effective punishment for foul play. He lived to see his idea reach fruition but then died, a year later, after a long illness. (Trivia alert: On September 14, 1891, the Wolverhampton Wanderers were awarded the first penalty kick in a football league in a game against Accrington Stanley. The penalty was taken and scored by "Billy" Heath as the Wolves went on to win the game 5–0.)

McCrum's legacy is enormous, considering the worldwide importance of soccer today and the significance of the penalty kick within the game. He would have, no doubt, been proud to see how central his idea became to the overall development of the game. However, not even in his wildest dreams could he have anticipated that his penalty kick could also provide the data necessary to verify for the first time a mathematical theorem that was fundamental in economics: the minimax theorem. This is the objective of this chapter.

A formal model of the penalty kick may be written as follows. Let the player's payoffs be the probabilities of success ("score" for the kicker and "no score" for the goalkeeper) in the penalty kick. The kicker wishes to maximize the expected probability of scoring, and the goalkeeper wishes to minimize it. Consider, for example, a simple 2×2 game-theoretical model of player's actions for the penalty kick and let π_{ij} denote the kicker's probabilities of scoring, where $i = \{L,R\}$ denotes the kicker's choice and $j = \{L,R\}$ the goalkeeper's choice, with L = left, R = right:

	L	R
L	π_{LL}	π_{LR}
R	π_{RL}	π_{RR}

Each penalty kick involves two players: a kicker and a goalkeeper. In the typical kick in professional leagues, the ball takes about 0.3 seconds

to travel the distance between the penalty mark and the goal line. This is less time than it takes for the goalkeeper to react and move to the possible paths of the ball. Hence, both kicker and goalkeeper must move simultaneously. Players have few strategies available, and their actions are observable. There are no second penalties in the event that a goal is not scored. The initial locations of both the ball and the goalkeeper are always the same: The ball is placed on the penalty mark, and the goalkeeper positions himself on the goal line, equidistant from the goalposts. The outcome is decided, in effect, immediately (roughly within 0.3 seconds) after players choose their strategies.

The clarity of the rules and the detailed structure of this simultaneous one-shot play capture the theoretical setting of a zero-sum game extremely well. In this sense, it presents notable advantages over other plays in professional sports and other real-world settings. In baseball, pitchers and batters have many actions available, and there are numerous possible outcomes. In cricket and tennis, possible outcomes are limited, but players also have many strategic choices available. Even in these sports, the direction of the serve or the pitch, its spin, and the initial location of the opponent are all important strategic choices that are hard to quantify. For instance, the position of the player returning a tennis serve or attempting to hit a baseball affects the choice of strategy by the server or the pitcher. A key additional difficulty is that a serve or a pitch is not a simultaneous (static) but a sequential (dynamic) game, in that the outcome of the play is typically not decided immediately. After a player serves or a pitcher throws, often there is subsequent strategic play that plays a crucial role in determining the final outcome. Each point in these situations is more like part of a dynamic game with learning, where each player plays what in economics is known as a multi-armed bandit problem at the start of the match.[2] As such, these situations deviate substantially from the theoretical postulates put forward here, and notable difficulties arise both in modeling nonsimultaneous situations theoretically and in observing all strategic choices in a given play.

The penalty kick game has a unique Nash equilibrium in mixed strategies when

2 In a dynamic game, there probably are spillovers from point to point, whereas in a standard repeated zero-sum game, especially if repeated infrequently, there are no such payoff spillovers. For instance, in tennis, having served to the left on the first serve (and say, faulted) may effectively be "practice" in a way that makes the server momentarily better than average at serving to the left again. If this effect is important, the probability that the next serve should be inside the line should increase. In other words, there should be negative serial correlation in the choice of serve strategies rather than, as will be shown later, the random play (no serial correlation) that is predicted by minimax. Consistent with this hypothesis, the results in Walker and Wooders (2001) confirm that tennis players switch serving strategies too often to be consistent with random play.

$$\pi_{LR} > \pi_{LL} < \pi_{RL},$$
$$\pi_{RL} > \pi_{RR} < \pi_{LR}$$

If the play in a penalty kick can be represented by this model, then equilibrium play requires each player to use a mixed strategy. In this case, the equilibrium yields two sharp testable predictions about the behavior of kickers and goalkeepers:

1. Success probabilities—the probability that a goal will be scored (not scored) for the kicker (goalkeeper)—should be the same across strategies for each player.

   Formally, let g_L denote the goalkeeper's probability of choosing left. This probability should be chosen so as to make the kicker's probability of success identical across strategies. That is, g_L should satisfy $pk_L = pk_R$, where

   $$pk_L = g_L \cdot \pi_{LL} + (1 - g_L) \cdot \pi_{LR}$$
   $$pk_R = g_L \cdot \pi_{RL} + (1 - g_L) \cdot \pi_{RR}$$

   Similarly, the kicker's probability of choosing left, k_L, should be chosen so as to make the goalkeeper's success probabilities identical across strategies, $pg_L = pg_R$, where

   $$pg_L = k_L \cdot (1 - \pi_{LL}) + (1 - k_L) \cdot (1 - \pi_{RL})$$
   $$pg_R = k_L \cdot (1 - \pi_{LR}) + (1 - k_L) \cdot (1 - \pi_{RR})$$

2. Each player's choices must be serially independent given constant payoffs across games (penalty kicks). That is, individuals must be concerned only with instantaneous payoffs, and intertemporal links between penalty kicks must be absent. Hence, players' choices must be independent draws from a random process. Therefore, they should not depend on one's own previous play, on the opponent's previous play, on their interaction, or on any other previous actions.

The intuition for these two testable hypotheses is the following. In a game of pure conflict (zero-sum), if it would be disadvantageous for you to let your opponent see your actual choice in advance, then you benefit by choosing at random from your available pure strategies. The proportions in your mix should be such that the opponent cannot exploit your choice by pursuing any particular pure strategy out of those available to him or her—that is, each player should get the same average payoff when he or she plays any of his or her pure strategies against his or her opponent's mixture.

In what follows, we test whether these two hypotheses can be rejected using classical hypothesis testing and real data. Incidentally, this reject–no reject dichotomy may be quite rigid in situations where the theory makes precise point predictions, as in the zero-sum game that we study.[3]

Data were collected on 9,017 penalty kicks during the period September 1995–June 2012 from professional games in Spain, Italy, England, and other countries. The data come from a number of TV programs, such as the *English Soccer League* in the United States, *Estudio Estadio* and *Canal+ Fútbol* in Spain, *Novantesimo Minuto* in Italy, *Sky Sports Football* in the United Kingdom, and others. These programs review all of the best plays in the professional games played every week, including all penalty kicks that take place in each game. The data include the names of the teams involved in the match, the date of the match, the names of the kicker and the goalkeeper for each penalty kick, the choices they take (left, center, or right), the time at which the penalty kick is shot, the score at that time, and the final score in the match. They also include the kicker's kicking leg (left or right) and the outcome of the shot (goal or no goal).[4] Around 80% of all observations come from league matches in England, Spain, and Italy.[5] Together with Germany, these leagues are considered to be the most important in the world.

There are two types of kickers, depending on their kicking legs: left-footed and right-footed. Most kickers in the sample are right-footed, as is the case in the population of soccer players and in the general population. These two types have different strong sides. Left-footed kickers shoot more often to the left-hand side of the goalkeeper than to the right-hand side, whereas right-footed kickers shoot more often to the right-hand side. Basic anatomical reasons explain these different strengths.

To deal with this difference, it makes sense to "normalize" the game and rename choices according to their "natural sides." In other words, given that the roles are reversed for right-footed kickers and left-footed kickers, it would be erroneous to treat the games associated with these different types of kickers as equal. For this reason, in the remainder of the chapter, players' choices are renamed according to the kickers' natural sides. Whatever the kicker's strong foot actually is, R denotes "kicker's natural side" and L denotes "kicker's nonnatural side." When the kicker is right-footed, the natural side R is the right-hand side of the goalkeeper, and when the kicker is left-footed, it is the left-hand side of the goalkeeper. This notation means, for instance, that a left-footed

3 O'Neill (1991) suggests for these cases an alternative that is much less rigid than the reject–no reject dichotomy: a Bayesian approach to hypothesis testing combined with a measure of closeness of the results to the predictions.

4 The outcome "no goal" includes saves made by the goalkeeper and penalties shot wide, to the goalpost, or to the crossbar by the kicker, each in separate categories.

5 The rest come from cup competitions (elimination tournaments that are simultaneously played during the annual leagues) and from international games.

kicker kicking to the goalkeeper's right is the same as a right-footed kicker kicking to the goalkeeper's left. Thus, the goalkeeper plays the same game when he or she faces a left-footed or a right-footed kicker, but the actions are simply identified differently. All that matters is whether the kicker and goalkeeper pick the kicker's strong side *R* or the kicker's weak side *L*. Payoffs are the same for the two kicker types up to the renaming of the actions. The same argument goes for goalkeepers. They tend to choose right more often than left when facing a right-footed kicker and left more often than right when facing a left-footed kicker, but the scoring rates are statistically identical when they face the two player types after the renaming of the actions.[6]

Table 1.1 shows the relative proportions of the different choices made by the kicker and the goalkeeper (Left (*L*), Center (*C*), or Right (*R*)), with the total number of observations in the second left-most column. The first letter refers to the choice made by the kicker and the second to the choice made by the goalkeeper, both from the viewpoint of the goalkeeper. For instance, "*RL*" means that the kicker chooses to kick to the right-hand side of the goalkeeper and the goalkeeper chooses to jump to his or her left. The right-most column shows the scoring rate for a given score difference. The term "score difference" is defined as the number of goals scored by the kicker's team minus the number of goals scored by the goalkeeper's team at the time the penalty is shot. For instance, a −1 means that the kicker's team was behind by one goal at the time of the penalty kick.

The strategy chosen by goalkeepers coincides with the strategy followed by kickers in about half of all penalty kicks in the data set. Most are *RR* (30.5%); 16.7% are *LL,* and 0.9% are *CC*. Kickers kick to the center relatively rarely (6.8% of all kicks), whereas goalkeepers appear to choose *C* even less often (3.5%), perhaps because they already cover part of the center with their legs when they choose *R* or *L*. The percentage of kicks where players' strategies do not coincide with each other is almost equally divided between *LR* (21.6%) and *RL* (21.7%).

A goal is scored in 80.07% of all penalty kicks. The scoring rate is close to 100% when the goalkeeper's choice does not coincide with the kicker's, and it is over 60% when it coincides. The average number of goals per match in the sample is 2.59. It is thus no surprise to observe that in most penalty kicks the score difference is 0, 1, or −1 at the time of the shot. For these score differences, the scoring rate is slightly greater in

6 See Palacios-Huerta (2003). This statistical identity can be shown using a regression framework. The null hypothesis that kicker's types are perfectly symmetric corresponds to a finding that kicker-type fixed effects are jointly insignificantly different from zero in explaining whether a goal was scored, including variables that describe the state of the soccer match at the time the penalty is shot as controls. The same holds for goalkeepers facing the different types.

Table 1.1. Distribution of Strategies and Scoring Rates in Percentage Terms

Score Difference	#Obs	*LL*	*LC*	*LR*	*CL*	*CC*	*CR*	*RL*	*RC*	*RR*	Scoring Rate
0	3701	16.5	1.2	21.5	4.0	1.4	3.5	20.6	1.2	30.1	81.5
1	1523	15.1	0.5	16.2	4.2	1.5	2.6	29.9	0.5	29.5	78.1
−1	2001	16.1	1.5	23.3	2.1	0.0	1.7	20.3	1.0	34.0	80.3
2	607	11.9	3.4	19.4	5.2	1.5	0.7	23.7	1.7	32.5	75.7
−2	744	20.1	1.5	27.6	3.7	0.0	2.5	16.1	0.3	28.2	78.5
Others	441	28.8	0.5	27.3	0.5	0.6	1.4	17.7	0.5	22.7	82.4
All	9017	16.7	1.7	21.6	3.4	0.9	2.5	21.7	0.9	30.5	80.07

tied matches (81.5%), followed by the rate in matches where the kicker's team is behind by one goal (80.3%), and then by the rate in matches where his or her team is ahead by one goal (78.1%).

Before we begin any formal test, it is worth examining the extent to which observed behavior appears to be close to the Nash equilibrium predictions. Players in the sample choose either R or L 96.3% of the time, kickers 93.2% of the time, and goalkeepers, 96.5%.[7] In what follows, we consider the choice C as if it was the same as the natural choices.[8] The typical penalty kick may then be described by the simple 2×2 model outlined earlier. Thus a penalty kick has a *unique* Nash equilibrium, and the equilibrium requires each player to use a mixed strategy. As mentioned already, equilibrium theory makes two testable predictions about the behavior of kickers and goalkeepers: (1) Winning probabilities should be the same across strategies for both players, and (2) each player's strategic choices must be serially independent.

For all players in the sample, the empirical scoring probabilities are the following:

	g_L	$1-g_L$
k_L	59.11	94.10
$1-k_L$	93.10	71.22

where, as indicated above, k_L and g_L denote the nonnatural sides. We can now compute the mixed strategy Nash equilibrium in this game (minimax frequencies) and compare it with the actual mixing probabilities observed in the sample (see figures 1.1 and 1.2). Interestingly, we find that observed aggregate behavior is virtually *identical* to the theoretical predictions:

	g_L	$1-g_L$	k_L	$1-k_L$
Nash Predicted Frequencies	40.23%	59.77%	38.47%	61.53%
Actual Frequencies	41.17%	58.83%	38.97%	61.03%

7 Chiappori et al. (2002) study the aggregate predictions of a zero-sum game, rather than individual player choices and pay close attention to the possibility that C is an available pure strategy. They conclude that the availability of C as an action is not an issue. Their findings are also substantiated in the data set used in this chapter. This evidence means that a penalty kick may be described as a two-action game.

8 Professional players basically consider strategy C and the strategy of playing their natural side as equivalent. The reason is that they typically kick with the interior side of their foot, which allows for greater control of the kick, by approaching the ball running from their nonnatural side. This phenomenon makes choosing C or their natural side equally difficult.

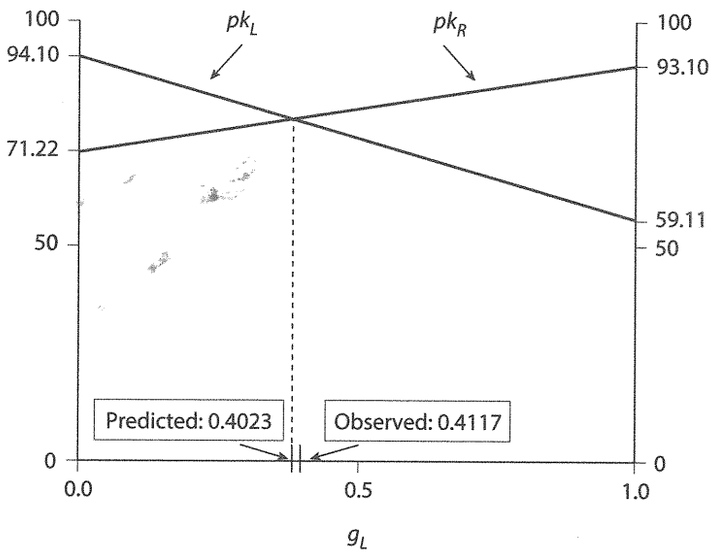


Figure 1.1. Nash and actual frequencies for goalkeepers.

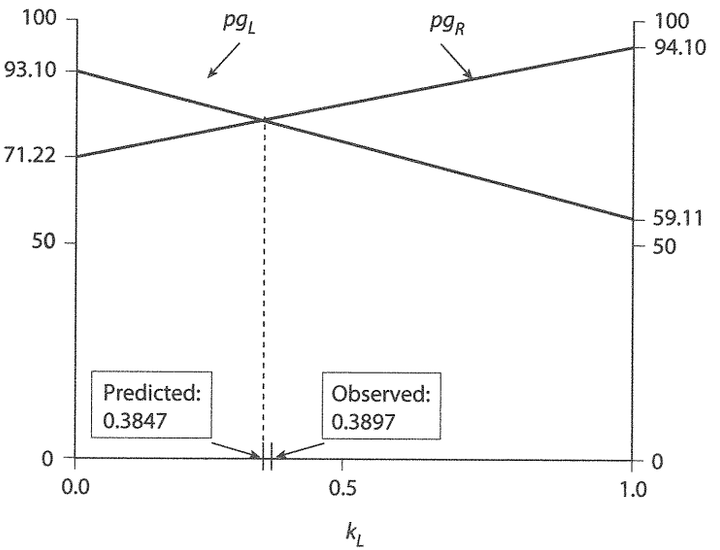


Figure 1.2. Nash and actual frequencies for kickers.

This is, at the very least, encouraging for the model. We turn next to testing the implications of the minimax theorem.

IMPLICATION NUMBER 1: TESTS OF EQUAL SCORING PROBABILITIES

The tests of the null hypothesis that the scoring probabilities for a player (kicker or goalkeeper) are identical across strategies can be implemented with the standard proportions tests, that is, using Pearson's χ^2 goodness-of-fit test of equality of two distributions.

Let p_{ij} denote the probability that player i will be successful when choosing strategy $j \in \{L,R\}$, n_{ij} the number of times that i chooses j, and N_{ijS} and N_{ijF} the number of times in which player i chooses strategy j and is successful (S) or fails (F) in the penalty kick. Success for a kicker is to score a goal, and for a goalkeeper is that a goal is not scored. Hence, we want to test the null hypothesis $p_{iL} = p_{iR} = p_i$. Statisticians tell us that to do this, the Pearson statistic for player i

$$p^i = \sum_{j \in \{L,R\}} \left[\frac{(N_{ijS} - n_{ij} p_i)^2}{n_{ij} p_i} + \frac{(N_{ijF} - n_{ij}(1 - p_i))^2}{n_{ij}(1 - p_i)} \right]$$

is distributed asymptotically as a χ^2 with 1 degree of freedom.

Quick statistical detour. In this and other statistical tests in this book, we will report p-values, so it is important to have a sense of what they are. Under the assumption that the hypothesis of interest (called the null hypothesis) is true, the p-value is the probability of obtaining a test statistic at least as extreme as the one that is actually observed. Thus, one often "rejects the null hypothesis" when the p-value is less than a predetermined significance level, often 0.05 (5%) or 0.01 (1%), indicating that the observed result would be highly unlikely under the null hypothesis. Many common statistical tests in this book, such as χ^2 tests or Student's t-test, produce test statistics that will be interpreted using p-values.

It is also possible to study whether behavior at the aggregate level is consistent with equilibrium play by testing the joint hypothesis that each individual case is *simultaneously* generated by equilibrium play. The test statistic for the Pearson joint test in this case is the sum of all the $\mathcal{N}$ individual test statistics, and under the null hypothesis this test is distributed as a χ^2 with $\mathcal{N}$ degrees of freedom. Note that this joint test allows for differences in probabilities p_i across players.

IMPLICATION NUMBER 2: TESTS OF RANDOMNESS OR SERIAL INDEPENDENCE

The second testable implication is that a player's mixed strategy is the same at each penalty kick. This notion implies that players' strategies are random or serially independent. Their play is not serially independent if, for instance, they choose not to switch their actions often enough or if they switch actions too often.

The work on randomization is extensive in the experimental economics and psychological literatures. Interestingly, this hypothesis has never found support in any empirical (natural and experimental) tests of the minimax hypothesis, and it is rarely supported in other tests. In particular, when subjects are asked to generate random sequences, their sequences typically have negative autocorrelation, that is, individuals exhibit a bias *against* repeating the same choice.[9] This phenomenon is often referred to as the "law of small numbers" (subjects may try to reproduce, in short sequences, what they know are the properties of long sequences). The only possible exception is Neuringer (1986), who explicitly taught subjects to choose randomly after hours of training by providing them with detailed feedback from previous blocks of responses in an experiment. These training data are interesting in that they suggest that experienced subjects might be able to learn to generate randomness. As Camerer (1995) remarks, "whether they do in other settings, under natural conditions, is an empirical question."

A simple way to test for randomness is to use the standard "runs test." Consider the sequence of strategies chosen by a player in the order in which they occurred $s = \{s_1, s_2, \ldots, s_n\}$ where $s_x \in \{L,R\}$, $x \in [1,n]$, and $n = n_R + n_L$ are the number of natural side and nonnatural side choices made by the player. Let r denote the total number of runs in the sequence s. A run is defined as a succession of one or more identical symbols that are followed and preceded by a different symbol or no symbol at all. Let $f(r, s)$ denote the probability that there are exactly r runs in the sequence s. Let $\Phi[r, s] = \Sigma_{k=1,\ldots,r} f(k,s)$ denote the probability of obtaining r or fewer runs. Gibbons and Chakraborti (1992) show that by using the exact mean and variance of the number of runs in an ordered sequence, then, under the null hypothesis that strategies are serially independent, the critical values for the rejection of the hypothesis can be found from the Normal distribution approximation to the null distribution.

More precisely, the variable

$$\frac{r - 0.5 - 2\left(\frac{n_L n_R}{n}\right)}{\left[2n_L n_R \left(\frac{2n_L n_R - n}{n^2(n-1)}\right)\right]^2}$$

9 See Bar-Hillel and Wagenaar (1991), Rapoport and Budescu (1992), Rapoport and Boebel (1992), and Mookherjee and Sopher (1994). Neuringer (2002), Rabin (2002) and Camerer (1995) review the literature. See also Tversky and Kahneman (1971).

is distributed as a standardized Normal probability distribution $N(0,1)$. The null hypothesis will then be rejected at the 5% confidence level if the probability of r or fewer runs is less than 0.025 or if the probability of r or more runs is less than 0.025, that is, if $\Phi[r, s] < 0.025$ or if $1 - \Phi[r-1, s] < 0.025$. Similarly, at the 10% level, the hypothesis is rejected if they are less than 0.05.[10]

The results in table 1.2 show the results of the Pearson test and the runs test for 40 world-class soccer players, half kickers, and half goalkeepers.

The null hypothesis of equality of payoffs cannot be rejected for the majority of players. It is rejected for just two players (David Villa and Frank Lampard) at the 5% significance level and four players at the 10% significance level (in addition to Villa and Lampard, Iker Casillas and Morgan De Sanctis). Note that we should expect some rejections, just as if we flip 40 coins 10 times each we should expect some coins, but not many, to yield by pure chance proportions that are far from 50–50, such as 9 heads and 1 tail, or 8 heads and 2 tails. The confidence levels we are willing to adopt (typically no greater than 5% or 10%) tell us how many rejections we should expect. In our case, with 40 players the expected number of rejections at the 5% level is $0.05 \times 40 = 2$, and at the 10% level, it is $0.10 \times 40 = 4$.

Thus, the evidence indicates that the hypothesis that scoring probabilities are identical across strategies cannot be rejected at the individual level for most players at conventional significance levels. The number of rejections is, in fact, identical to the theoretical predictions.

Furthermore, behavior at the aggregate level also appears to be consistent with equilibrium play. As already indicated, the joint hypothesis that each case is simultaneously generated by equilibrium play can be tested by computing the aggregate Pearson statistic (summing up the individual Pearson statistics) and checking if it is distributed as a χ^2 with N degrees of freedom. The results show that the Pearson statistic is 36.535 and its associated p-value is 0.627 for all 40 players. Hence, the hypothesis of equality of winning probabilities cannot be rejected at the aggregate level. Focusing only on kickers, the relevant statistic is 20.96 with a p-value of 0.399, and for goalkeepers it is 15.58 with a p-value of 0.742. Hence, the hypothesis of equality of winning probabilities cannot be rejected for either subgroup.

With respect to the null hypothesis of randomness, the runs tests show that this hypothesis cannot be rejected for the majority of players. They neither appear to switch strategies too often or too infrequently,

10 Aggregate level tests may also be implemented by checking if the values in columns $\Phi[r, s]$ and $\Phi[r-1, s]$ tend to be uniformly distributed in the interval [0, 1], which is what should happen under the null hypothesis of randomization. See Palacios-Huerta (2003).

but just about the right amount. This hypothesis is in fact rejected for just three players (David Villa, Alvaro Negredo, and Edwin Van der Sar), and four players (in addition, Jens Lehman) at the 5% and 10% significance levels. For the same reasons as in the previous test, we should be expecting two and four rejections.

The runs test is simple and intuitive. However, it is a test that has low power to identify a lack of randomness. Put differently, current choices may be explained, at least in part, by past variables such as past choices or past outcomes, or past choices of the opponent, or interactions with these variables, and still the number of runs in the series of choices may appear to be neither too high nor too low. As such, many potential sources of dynamic dependence cannot be detected with a runs test. For this reason, some researchers on randomization have studied whether past choices or outcomes have any role in determining current choices by estimating a logit equation for each player. For instance, in Brown and Rosenthal (1990), the dependent variable is a dichotomous indicator of the current choice of strategy, and the independent variables are first and second lagged indicators for both players' past choices, the products of their first and second lagged choices, and an indicator for the opponent's current choices. The results show that in fact it is possible to detect a number of dynamic dependences with this logit equation that are not possible to detect with the runs test.[11]

Unfortunately, the standard logit equation is still problematic in that the way this procedure is typically implemented generates *biased* estimates. We will take a quick technical detour to explain why. The choice of strategy in a penalty kick may depend on certain observed characteristics of the player and his or her opponent, the specific sequence of past choices and past outcomes, and perhaps other variables. It may also depend on unobserved characteristics. Thus, the basic econometric problem is to estimate a binary choice model with lagged endogenous variables and unobserved heterogeneity where the effect of state dependence needs to be controlled for appropriately. The econometric estimation of these models is subject to a number of technical difficulties. For example, parameter estimates jointly estimated with individual fixed effects can be seriously *biased* and *inconsistent*. Arellano and Honoré (2001) offer an excellent review.

To establish the idea that past choices have no significant role in determining current choices, we estimate a logit equation for each

11 Compare table IV in Brown and Rosenthal (1990) with table 4 in Walker and Wooders (2001). There are many subjects that pass the runs test but still exhibit serial dependence in that a number of lagged endogenous variables (choices and outcomes) help predict their subsequent choices.

Table 1.2. Pearson and Runs Tests

		Proportions		Success Rate		Pearson Tests		Runs Tests		
Name	#Obs	L	R	L	R	Statistic	p-value	r	$\Phi[r-1, s]$	$\Phi[r, s]$
Kickers:										
Mikel Arteta	53	0.433	0.566	0.782	0.833	0.218	0.639	27	0.439	0.551
Alessandro Del Piero	55	0.345	0.654	0.736	0.805	0.344	0.557	24	0.237	0.339
Samuel E'too	62	0.419	0.580	0.769	0.805	0.120	0.728	28	0.165	0.239
Diego Forlán	62	0.419	0.580	0.769	0.805	0.120	0.728	30	0.327	0.427
Steven Gerrard	50	0.340	0.660	0.823	0.909	0.777	0.377	23	0.382	0.507
Thierry Henry	44	0.477	0.522	0.809	0.782	0.048	0.825	19	0.086	0.145
Robbie Keane	42	0.309	0.690	0.769	0.758	1.174	0.278	17	0.184	0.296
Frank Lampard	38	0.236	0.763	0.666	0.793	4.113	0.042**	17	0.791	0.898
Lionel Messi	45	0.377	0.622	1.000	0.928	1.270	0.259	22	0.416	0.544
Alvaro Negredo	45	0.288	0.711	0.769	0.906	1.501	0.220	26	0.986**	0.995
Martín Palermo	55	0.381	0.618	0.714	0.735	0.028	0.865	23	0.098	0.158
Andrea Pirlo	39	0.384	0.615	0.733	0.833	0.566	0.451	20	0.505	0.639
Xabi Prieto	37	0.324	0.675	0.833	0.880	0.151	0.697	16	0.256	0.392
Franc Ribéry	38	0.500	0.500	0.789	0.736	0.145	0.702	25	0.930	0.964
Ronaldinho	46	0.456	0.543	0.952	0.880	0.753	0.385	24	0.460	0.580
Christiano Ronaldo	51	0.372	0.627	0.842	0.718	1.008	0.315	24	0.342	0.458
Roberto Soldado	40	0.400	0.600	0.937	0.750	2.337	0.126	21	0.539	0.667
Francesco Totti	47	0.489	0.510	0.782	0.833	0.195	0.658	20	0.070	0.119
David Villa	52	0.365	0.634	0.631	0.909	5.978	0.014**	18	0.010	0.022**
Zinedine Zidane	61	0.377	0.622	0.782	0.815	0.099	0.752	26	0.126	0.192
All	962	0.386	0.613	0.795	0.822	20.96	0.399			
Goalkeepers:										
Dani Aranzubia	68	0.455	0.544	0.225	0.189	0.138	0.709	29	0.062	0.098
Gianluigi Buffon	71	0.408	0.591	0.241	0.142	1.113	0.291	35	0.420	0.518
Willie Caballero	60	0.350	0.650	0.095	0.230	1.674	0.195	29	0.522	0.634
Iker Casillas	69	0.347	0.652	0.250	0.088	3.278	0.070*	32	0.414	0.520
Petr Čech	82	0.414	0.585	0.235	0.187	0.276	0.590	38	0.224	0.298
Júlio César	68	0.308	0.691	0.238	0.106	2.007	0.156	34	0.840	0.900
Morgan De Sanctis	62	0.435	0.564	0.148	0.342	3.018	0.082*	34	0.700	0.783
Tim Howard	67	0.402	0.597	0.222	0.225	0.000	0.978	30	0.169	0.241
Bodo Illgner	68	0.352	0.647	0.250	0.272	0.041	0.839	33	0.547	0.650
Gorka Iraizoz	73	0.424	0.575	0.129	0.142	0.028	0.865	32	0.106	0.157
David James	69	0.391	0.608	0.185	0.238	0.270	0.603	40	0.924	0.954
Oliver Kahn	58	0.379	0.620	0.227	0.138	0.747	0.387	33	0.881	0.928
Andreas Kopke	70	0.428	0.571	0.233	0.150	0.787	0.374	31	0.119	0.175
Jens Lehman	72	0.444	0.555	0.218	0.225	0.004	0.949	28	0.014	0.026*
Andrés Palop	66	0.439	0.560	0.206	0.297	0.694	0.404	34	0.498	0.597
Pepe Reina	55	0.418	0.581	0.173	0.187	0.016	0.897	31	0.778	0.852
Mark Schwarzer	55	0.381	0.618	0.238	0.264	0.048	0.825	31	0.846	0.904
Stefano Sorrentino	48	0.458	0.541	0.136	0.269	1.275	0.258	27	0.687	0.783
Víctor Valdes	71	0.394	0.605	0.214	0.232	0.032	0.857	32	0.196	0.272
Edwin van der Sar	80	0.412	0.587	0.121	0.148	0.125	0.722	26	0.000	0.001**
All	1332	0.402	0.597	0.199	0.198	15.58	0.742			

Notes: ** and * denote rejections at the 5% and 10% levels, respectively.

Table 1.3. Results of Significance Tests (Logit) for the Choice of the Natural Side

Null Hypothesis		Players Whose Behavior Allows Rejection of the Null Hypothesis at the: 5% Level	10% Level
A. $a_1 = a_2 = b_0 = b_1 = b_2 = c_1 = c_2 = 0$	Kicker	—	David Villa, Frank Lampard
	Goalkeeper	—	Iker Casillas
B. $a_1 = a_2 = 0$	Kicker	—	David Villa
	Goalkeeper	—	Andreas Kopke
C. $b_1 = b_2 = 0$	Kicker	—	—
	Goalkeeper	—	Jens Lehman
D. $c_1 = c_2 = 0$	Kicker	—	Martín Palermo
	Goalkeeper	—	—
E. $b_0 = 0$	Kicker	—	Ronaldinho
	Goalkeeper	—	Júlio César, Edwin van der Sar

Here is the estimating equation:

$$R = G[a_0 + a_1\text{lag}(R) + a_2\text{lag2}(R) + b_0R^* + b_1\text{lag}(R^*) + b_2\text{lag2}(R^*) + c_1\text{lag}(R)\text{lag}(R^*) + c_2\text{lag2}(R)\text{lag2}(R^*)]$$

Notes: The asterisk * denotes the choice of the opponent. The terms "lag" and "lag2" denote the choices previously followed in the ordered sequence of penalty kicks. The function $G[x]$ denotes $\exp(x)/[1+\exp(x)]$.

player based on the Arellano and Carrasco (2003) method using the same specification as Brown and Rosenthal (1990). The model generates *unbiased* and *consistent* estimates and it allows for unobserved heterogeneity and for individual effects to be correlated with the explanatory variables (see table 1.3).

The main finding is that the null hypothesis of randomization (implication number 2), that all the explanatory variables are jointly statistically insignificant (hypothesis A), cannot be rejected for any player at the 5% level and is rejected for only three players (David Villa, Frank Lampard, and Iker Casillas) at the 10% level.

The table also reports the tests of different subhypotheses concerning whether one's past choices alone, or past opponent's choices alone, or successful past plays alone may determine current choices. No evidence that any player made choices in a serially dependent fashion in any respect is found at the 5% level, and at the 10% level, none of the hypotheses are rejected for more than two players. These results indicate that

the choices of most players are unrelated to their own previous choices and outcomes and to their opponents' previous choices and outcomes, exactly as in a random series.

A number of extensions of this investigation are possible. From a more technical perspective, for instance, the statistical power of the tests in various ways, as well as the ability of the tests to detect deviations from minimax play, can be studied using Monte Carlo simulations. From a more empirical perspective, we may consider more strategies such as *C* and others, and then test the implications in a 3×3 game or in an $N \times N$ game rather than in a 2×2 game.[12]

The main finding in this chapter is that the results of the tests are remarkably consistent with equilibrium play in every respect: (1) Winning probabilities are statistically identical across strategies for players, and (2) players generate serially independent sequences and ignore possible strategic links between subsequent penalty kicks. These results, which extend Palacios-Huerta (2003), represent the first time that both implications of von Neumann's (1928) minimax theorem are supported in real life.

*

In recent years, the tests in this chapter have been used to advise a number of teams participating in some of the main club tournaments in the world (e.g., UEFA Champions League and the Football Association Challenge Cup in England, known as the FA Cup), as well as some national teams participating in the top event in the world taking place every four years: the World Cup (Kuper 2011).

In particular, these tests were first used in the UEFA Champions League final on May 21, 2008, in Moscow, to advise Chelsea in its penalty shoot-out versus Manchester United. At the time, no one in the media noticed a pattern in the behavior of the players in the shoot-out, not even a number of small but critical incidents. No one understood what the players were doing and why they were doing it. There was no model to make sense of any behavior. The story is described in great detail in *Soccernomics* (2012) by Simon Kuper and Stefan Szymanski, and this is not the place to repeat it entirely. But it is perhaps worth quoting a few sentences:

> So far, the advice [of the tests] had worked very well for Chelsea (The right-footed penalty-takers had obeyed it to the letter, Manchester United's goalkeeper Van der Sar had not saved a single penalty, and

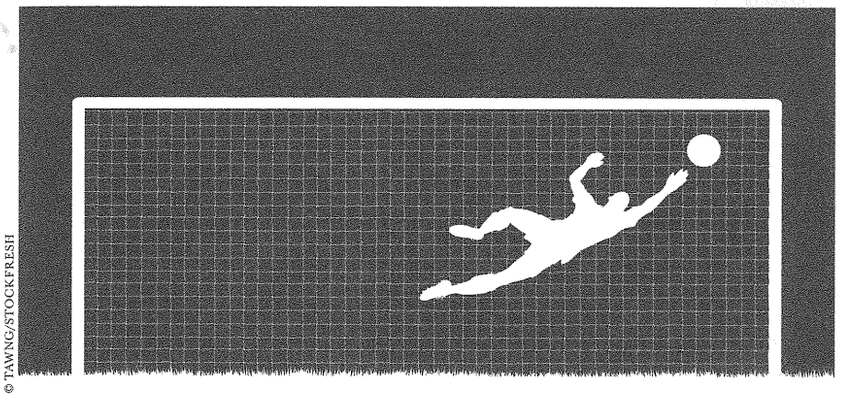
12 See the appendix in Palacios-Huerta (2003) with some evidence on the 3×3 game.

Chelsea's keeper had saved Cristiano Ronaldo's) . . . As Anelka prepared to take Chelsea's seventh penalty, the gangling keeper, standing on the goal-line, extended his arms to either side of him. Then, in what must have been a chilling moment for Anelka, the Dutchman [Van der Sar] pointed with his left hand to the left corner. "That's where you're all putting it, isn't it?" he seemed to be saying. Now Anelka had a terrible dilemma. This was game theory in its rawest form. . . . So Anelka knew that Van der Sar knew that Anelka knew that Van der Sar tended to dive right against right-footers. What was Anelka to do?

You may perhaps know the end. If you do not, this is the authors' summary: "Anelka's decision to ignore the advice [of these tests] probably cost Chelsea the Champions League."

2

VERNON SMITH MEETS MESSI IN THE LABORATORY



© TAWNG/STOCKFRESH

When the exact question being asked and the population being studied are mirrored in a laboratory experiment, the information from the experiment can be clear and informative.
—Armin Falk and James Heckman 2009

A few years ago, a match in the Argentine provinces had to be abandoned just seconds before the finish when the referee, who had just awarded a penalty, was knocked out by an irate player. The league court decided that the last 20 seconds of the game—the penalty kick, in effect—should be played the next Sunday. That gave everyone a week to prepare for the penalty.

At dinner a few nights before the penalty, the goalkeeper, El Gato Díaz, mused about the kicker: "Messi kicks to the right."

"Always," said the president of the club.

"But Messi knows that I know."

"Then, we are fucked."

"Yeah, but I know that he knows," said El Gato.

"Then dive to the left and be ready," said someone at the table.