

Vol. 31, No. 3, 2018

# CHANCE

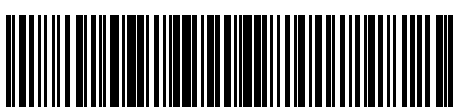
Using Data to Advance Science, Education, and Society

## Special Issue on **Sports**

### **Including...**

**Luck and Skill in  
Tournament Golf**

**All in the Family:  
German Twins' Finishing  
Times in the 2016 Women's  
Olympic Marathon**



09332480(2018)31(3)



**Taylor & Francis**  
Taylor & Francis Group

**ASA**

# EXCLUSIVE BENEFITS FOR ALL ASA MEMBERS!

**SAVE 30%** on Book Purchases with discount code **ASA18**.

Visit the new ASA Membership page to unlock savings on the latest books,  
access exclusive content and review our latest journal articles!

With a growing recognition of the importance of statistical reasoning across many different aspects of everyday life and in our data-rich world, the American Statistical Society and CRC Press have partnered to develop the **ASA-CRC Series on Statistical Reasoning in Science and Society**. This exciting book series features:

- Concepts presented while assuming minimal background in Mathematics and Statistics.
- A broad audience including professionals across many fields, the general public and courses in high schools and colleges.
- Topics include Statistics in wide-ranging aspects of professional and everyday life, including the media, science, health, society, politics, law, education, sports, finance, climate, and national security.

## DATA VISUALIZATION

Charts, Maps, and Interactive Graphics

**Robert Grant**, BayersCamp

This book provides an introduction to the general principles of data visualization, with a focus on practical considerations for people who want to understand them or start making their own. It does not cover tools, which are varied and constantly changing, but focusses on the thought process of choosing the right format and design to best serve the data and the message.

September 2018 • 210 pp • Pb: 9781138707603: \$29.95 \$23.96 • [www.crcpress.com/9781138707603](http://www.crcpress.com/9781138707603)

## VISUALIZING BASEBALL

**Jim Albert**, Bowling Green State University, Ohio, USA

A collection of graphs will be used to explore the game of baseball. Graphical displays are used to show how measures of batting and pitching performance have changed over time, to explore the career trajectories of players, to understand the importance of the pitch count, and to see the patterns of speed, movement, and location of different types of pitches.

August 2017 • 142 pp • Pb: 9781498782753: \$29.95 \$23.96 • [www.crcpress.com/9781498782753](http://www.crcpress.com/9781498782753)

## ERRORS, BLUNDERS, AND LIES

How to Tell the Difference

**David S. Salsburg**, Emeritus, Yale University, New Haven, CT, USA

In this follow-up to the author's bestselling classic, "The Lady Tasting Tea", David Salsburg takes a fresh and insightful look at the history of statistical development by examining errors, blunders and outright lies in many different models taken from a variety of fields.

April 2017 • 154 pp • Pb: 9781498795784: \$29.95 \$23.96 • [www.crcpress.com/9781498795784](http://www.crcpress.com/9781498795784)



JOURNAL OF THE AMERICAN  
STATISTICAL ASSOCIATION  
Vol 112, 2017

THE AMERICAN STATISTICIAN  
Vol 72, 2018

STATISTICS AND PUBLIC POLICY  
Vol 5, 2018



Taylor & Francis Group  
an informa business

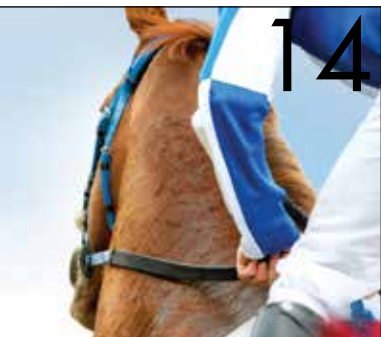


<http://bit.ly/CRCASA2018>

# CHANCE

Using Data to Advance Science, Education, and Society

<http://chance.amstat.org>



## ARTICLES

- 4 Luck and Skill in Tournament Golf  
*Stephen M. Stigler and Margaret L. Stigler*
- 14 Why a 1-for-45 Record in the Kentucky Derby Does Not Necessarily Equate to Underachievement  
*Leonard Cupingood*
- 20 All in the Family: German Twins' Finishing Times in the 2016 Women's Olympic Marathon  
*David Cottrell and Michael C. Herron*
- 29 The Point(s)-After-Touchdown Decision Revisited  
*Harold Sackrowitz*
- 37 An Analysis of the First Round of the MLB First-Year Player Draft  
*Gabriel Chandler and Simon Rosenbaum*
- 44 An Ordinal Logistic Regression Model for the Masters Golf Tournament  
*Erik L. Heiny and Cody C. Frisby*
- 59 Defying the Odds: How Likely Are We to See Another Team Pull a 'Leicester' and Win the EPL?  
*Craig A. Heard and A. John Bailer*

## COLUMNS

### 67 The Odds of Justice

Mary W. Gray, Column Editor

Code of Silence

*How private companies hide flaws in the software that governments use to decide who goes to prison and who gets out*

*Rebecca Wexler*

## DEPARTMENTS

- 3 Editor's Letter

---

*Abstracted/indexed in Academic OneFile, Academic Search, ASFA, CSA/Proquest, Current Abstracts, Current Index to Statistics, Gale, Google Scholar, MathEDUC, Mathematical Reviews, OCLC, Summon by Serial Solutions, TOC Premier, Zentralblatt Math.*

*Cover image: Getty Images*



## EXECUTIVE EDITOR

**Scott Evans**

Harvard School of Public Health, Boston, Massachusetts  
[evans@sdac.harvard.edu](mailto:evans@sdac.harvard.edu)

## ADVISORY EDITORS

**Sam Behseta**

California State University, Fullerton

**Michael Larsen**

St. Michael's College, Colchester, Vermont

**Michael Lavine**

University of Massachusetts, Amherst

**Dalene Stangl**

Carnegie Mellon University, Pittsburgh, Pennsylvania

**Hal S. Stern**

University of California, Irvine

## EDITORS

**Jim Albert**

Bowling Green State University, Ohio

**Phil Everson**

Swarthmore College, Pennsylvania

**Dean Follman**

NIAID and Biostatistics Research Branch, Maryland

**Toshimitsu Hamasaki**

Office of Biostatistics and Data Management  
National Cerebral and Cardiovascular Research  
Center, Osaka, Japan

**Jo Hardin**

Pomona College, Claremont, California

**Tom Lane**

MathWorks, Natick, Massachusetts

**Michael P. McDermott**

University of Rochester Medical Center, New York

**Mary Meyer**

Colorado State University at Fort Collins

**Kary Myers**

Los Alamos National Laboratory, New Mexico

**Babak Shahbaba**

University of California, Irvine

**Lu Tian**

Stanford University, California

## COLUMN EDITORS

**Di Cook**

Iowa State University, Ames  
*Visiphilia*

**Chris Franklin**

University of Georgia, Athens  
*K-12 Education*

**Andrew Gelman**

Columbia University, New York, New York  
*Ethics and Statistics*

**Mary Gray**

American University, Washington, D.C.  
*The Odds of Justice*

**Shane Jensen**

Wharton School at the University of Pennsylvania,  
Philadelphia  
*A Statistician Reads the Sports Pages*

**Nicole Lazar**

University of Georgia, Athens  
*The Big Picture*

**Bob Oster**, University of Alabama, Birmingham, and  
**Ed Gracely**, Drexel University, Philadelphia, Pennsylvania  
*Teaching Statistics in the Health Sciences*

**Christian Robert**

Université Paris-Dauphine, France  
*Book Reviews*

**Aleksandra Slavkovic**

Penn State University, University Park  
*O Privacy, Where Art Thou?*

**Dalene Stangl**, Carnegie Mellon University, Pittsburgh,  
Pennsylvania, and **Mine Çetinkaya-Rundel**,  
Duke University, Durham, North Carolina  
*Taking a Chance in the Classroom*

**Howard Wainer**

National Board of Medical Examiners, Philadelphia,  
Pennsylvania  
*Visual Revelations*

## WEBSITE

<http://chance.amstat.org>

## AIMS AND SCOPE

*CHANCE* is designed for anyone who has an interest in using data to advance science, education, and society. *CHANCE* is a non-technical magazine highlighting applications that demonstrate sound statistical practice. *CHANCE* represents a cultural record of an evolving field, intended to entertain as well as inform.

## SUBSCRIPTION INFORMATION

*CHANCE* (ISSN: 0933-2480) is co-published quarterly in February, April, September, and November for a total of four issues per year by the American Statistical Association, 732 North Washington Street, Alexandria, VA 22314, USA, and Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA.

**U.S. Postmaster:** Please send address changes to *CHANCE*, Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA.

## ASA MEMBER SUBSCRIPTION RATES

ASA members who wish to subscribe to *CHANCE* should go to ASA Members Only, [www.amstat.org/membersonly](http://www.amstat.org/membersonly) and select the "My Account" tab and then "Add a Publication." ASA members' publications period will correspond with their membership cycle.

## SUBSCRIPTION OFFICES

**USA/North America:** Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA. Telephone: 215-625-8900; Fax: 215-207-0050. UK/Europe: Taylor & Francis Customer Service, Sheepen Place, Colchester, Essex, CO3 3LP, United Kingdom. Telephone: +44-(0)-20-7017-5544; fax: +44-(0)-20-7017-5198.

For information and subscription rates please email [subscriptions@tandf.co.uk](mailto:subscriptions@tandf.co.uk) or visit [www.tandfonline.com/pricing/journal/ucha](http://www.tandfonline.com/pricing/journal/ucha).

## OFFICE OF PUBLICATION

American Statistical Association, 732 North Washington Street, Alexandria, VA 22314, USA. Telephone: (703) 684-1221. Editorial Production: Megan Murphy, Communications Manager; Valerie Nirala, Publications Coordinator; Ruth E. Thaler-Carter, Copyeditor; Melissa Gotherman, Graphic Designer. Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA. Telephone: (215) 625-8900; Fax: (215) 207-0047.

Copyright ©2018 American Statistical Association. All rights reserved. No part of this publication may be reproduced, stored, transmitted, or disseminated in any form or by any means without prior written permission from the American Statistical Association. The American Statistical Association grants authorization for individuals to photocopy copyrighted material for private research use on the sole basis that requests for such use are referred directly to the requester's local Reproduction Rights Organization (RRO), such as the Copyright Clearance Center ([www.copyright.com](http://www.copyright.com)) in the United States or The Copyright Licensing Agency ([www.cla.co.uk](http://www.cla.co.uk)) in the United Kingdom. This authorization does not extend to any other kind of copying by any means, in any form, and for any purpose other than private research use. The publisher assumes no responsibility for any statements of fact or opinion expressed in the published papers. The appearance of advertising in this journal does not constitute an endorsement or approval by the publisher, the editor, or the editorial board of the quality or value of the product advertised or of the claims made for it by its manufacturer.

## RESPONSIBLE FOR ADVERTISEMENTS

For advertising inquiries, please contact [advertising@taylorandfrancis.com](mailto:advertising@taylorandfrancis.com). Printed in the United States on acid-free paper.



Scott Evans

## Dear CHANCE Colleagues,

Children across the globe are fascinated with sports statistics. For many, this fascination never goes away! That's why we devote this special themed issue of *CHANCE* to sports.

Interest in sports statistics has grown rapidly over the past decade. We have the data! My friend and colleague, Mark Glickman, and I founded the New England Symposium on Statistics in Sports (NESSIS) at Harvard University. In its first offering in 2007, we had 110 registrants. Since then, we have held NESSIS biannually with growth to 125, 150, 180, 240, and 245 registrants in 2009, 2011, 2013, 2015, and 2017 respectively. (Please consider joining us for NESSIS VII in September 2019.)

This is an exciting time for statistics in sports. New technologies have been created and an explosion of data is available for analyses. The "Moneyball" culture has spread to many other sports beyond baseball. It is difficult to identify professional teams that do not have a developing group of statistical analysts to evaluate players and game strategies.

In this special issue, we have eight interesting articles with statistics applications to baseball, football, golf, horse racing, marathon running, and soccer.

The honors go to the Stiglers. **Stephen** and **Margaret Stigler** use data from the four major men's golf tournaments from 1994–2015 and two of the four major women's golf tournaments from 10 unspecified years to determine how much of the variation in scores is due to skill level (the persistent capacity to play at a certain level) and how much is due to "luck" (transient variations in a player's score). Their analysis

leads to interesting insights regarding the relative contributions of skill level and luck that depend on the particular major tournament for men and the sex of the players.

In other articles, **Leonard Cupingood** illustrates why a 1-for-45 record in the Kentucky Derby does not imply underachievement. **David Cottrell** and **Michael Herron** evaluate the curious case of German twins Anna and Lisa Hahner, who finished the 2016 Olympic marathon simultaneously. Was this intentional or coincidental? **Harold Sackrowitz** revisits the point(s)-after-touchdown decision, three years after the NFL made the rule change that the one-point kick conversion was to be made from 33 rather than 20 yards. **Gabriel Chandler** and **Simon Rosenbaum** analyze the first round of the MLB First-Year Player Draft. **Erik Heiny** and **Cody Frisby** describe an ordinal logistic regression model that can be used to estimate score probabilities for each player and hole in the Masters. **Craig Heard** and **A. John Bailer** evaluate the likelihood of another Leicester-type miracle in the EPL.

Beyond sports, in the column "The Odds of Justice," **Rebecca Wexler** discusses a code of silence and how companies hide flaws in software that is used by governments for imprisonment and release decisions.

Overall, this is an exciting issue for fans of almost any sport. We hope you enjoy it.

*Scott Evans*

# Luck and Skill in Tournament Golf

*Stephen M. Stigler and Margaret L. Stigler*

On June 16, 2005, a field of 156 of the finest male golfers in the world played the first round of the 2005 U.S. Open, one of the four “major” annual tournaments in men’s professional golf. The list of entrants included nearly every major tournament player of the past few years, and all of the players were present either by invitation or because of excellent play in previous outings. Most of them were players at or near the top of their game, and had to be considered threats to win or at least place highly in this rich tournament.

With a prize of over \$1 million to the winner, all could be expected to give that initial round their maximum concentration—yet, they played that round with widely varying results: The average score for the day was 75, and the range ran from 67 to 85.

All 156 of these competitors were splendid golfers; 148 were top professionals and the other eight were the best amateurs in the country. Why should their performances vary so much on a single day?

Two reasons present themselves. The first is that while all were splendid golfers, perhaps they were not equally splendid: There was, without doubt, some variation in skill levels. The second is that the entrants benefited unequally from what is commonly referred to as “luck.”

An examination of the sizes and relative importance of these two components of the variation in scores of top tournament golfers is a principal goal of the study discussed here. A secondary and more-speculative aim is to see what implications can be drawn from this narrowly focused and quantitative study about the balance of skill and luck more generally—in other sports and in the social world.

## Skill and Luck

“Skill” and “luck” are common terms in everyday language, but the sense in which they are used here requires discussion.

“Luck” is perhaps the simplest of the two: By a player’s luck, we refer to transient variations in a player’s score; that is, variations particular to that player (which do not reflect some common cause that affects all players that day). Luck does not tend to persist from day to day over the four days of the tournament. Luck may be score variations due to bad or good bounces or wind gusts, to chance encounters with obstructions or spectators, to nervous reactions of a moment.

We specifically do not count as luck the good fortune that seems to belong to certain players



systematically: Such luck is persistent and properly called skill. By luck, we mean changes in score that would occur for a single player if he or she were to replay the same round under exactly the same conditions, day after day, without any benefit from the experience.

“Skill” is more difficult to describe, particularly because we mean to emphasize aspects of skill that differ from ordinary understanding. All of the entrants possess skills at the very highest level; it would seem foolish to raise even the slightest doubt that skill is the predominant determinant of success in tournament golf, but a little reflection shows the error of this way of thinking.

Suppose for a moment that the very best golfer in the world could be cloned, and imagine a tournament consisting of 156 exact clones of that golfer. Skill would play no role at all in the outcome of that imaginary tournament—it would be decided entirely by what we call luck. Yet if we were to add a single ordinary mortal to that pool of talent, the situation would change dramatically, and the skill difference of the mortal and the clones would be evident to all.

The point is that the absolute level of skill of the players is unimportant to the result of any tournament. Only *variation* in the skill levels of all those competing is important; only the relative skills of those competing play a role in the determination of the winner.

Skill in each golfer is the persistent capacity to play at a certain level, a capacity that we will suppose remains unchanged over the four days of the tournament. If a single player were to play the same course under exactly the same conditions repeatedly for many days, then the difference between that player’s average score per round and the similar average for all tournament entrants will be that player’s skill.

Our principal object is the study of the variation in different players’ skills among the entrants and the comparative magnitudes of this variation and the variation in luck for individual players. This relationship will influence the outcome of the tournament. When the skill variation dominates, the more-skillful players will finish near the top consistently. When luck variation dominates, no single player or group of players will win consistently.

Since a tournament extends over four days, we recognize one other sort of variation: the day-to-day variation due to conditions supposed to affect all players equally, such as pin placement, weather condition, or general media stress. We comment later on the possible effect such variation may have on the analysis if it affects only some of the players, such as a weather change in the middle of the day.

Major golf tournaments (other than “match play”) consist of 72 holes of play (each with a specified “par,” a measure of difficulty equal to the number of strokes expected to be taken to complete the hole; generally 3, 4, or 5). The play is spread out over four days, with 18 holes played each day.

The winner is the golfer who completes all 72 holes with the lowest score; ties are usually decided by the tied players playing extra holes until the tie is broken. A large number of players (often 120 to 160) starts the tournament. At the end of the second day, when 36 holes have been completed, about half the players are “cut” and sent home. A very few of the entrants leaves because of injury or for personal reasons.

The difficulty of the course can be affected by changes in weather or by the fact that placement of the pins is changed after each 18-hole round. If weather causes a long interruption in play, the unfinished rounds may be completed the next day, before the next round begins. In “match play” (not considered here), each day’s round is decided by the number of holes won, not the strokes taken.

## A Simple Model

Tournament golf is a very complicated enterprise, so it is surprising that a great deal can be learned from a model that ignores much of the complication. The proposed model is the essence of simplicity:

$$\text{Player's score in a round} = \text{par for the day} + \text{player's skill} + \text{luck}.$$

More formally, the score of the  $i$ th player on the  $j$ th day of the four day tournament,  $X_{ij}$ , can to a reasonable approximation be given as:

$$X_{ij} = \mu_j + S_i + L_{ij}.$$

Here  $\mu_j$  is the par for the day: notionally, if these same players were to play repeatedly exactly the same course under the same conditions, this would be their average score. The skill of the  $i$ th player,  $S_i$ , would be the average difference between that player’s score and the daily par (average of  $X_{ij} - \mu_j$ ) over a long run of play where the player’s ability remains unchanged.

Note that high scores are not desirable in golf, so a large positive  $S_i$  corresponds to low skill here. Players with high skill would have negative  $S_i$ .  $L_{ij}$  is the player’s luck in that round, defined as:

$$L_{ij} = X_{ij} - \mu_j - S_i.$$



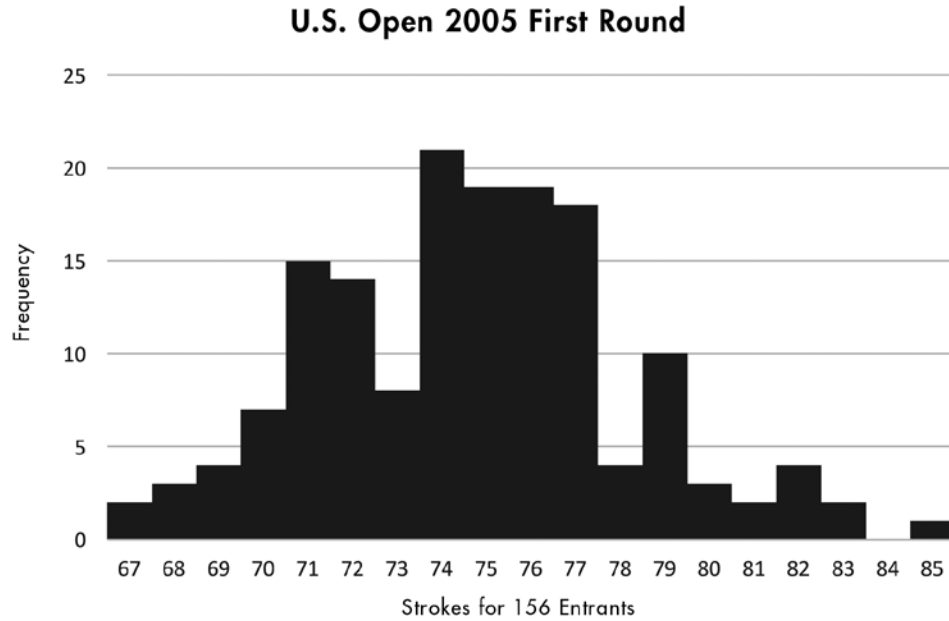


Figure 1. Results of first round of U.S. Open, June 16, 2005.

Expressed that way, the relationship is tautological, but it ceases to be so with a few further assumptions: We suppose, most importantly, that the luck and skill are statistically independent—that players at all levels are subject to the same magnitude in variation due to luck, be they average tournament players, the top rung in the tournament, or the below-average. That is, the level of play varies with the skill  $S$ , but variation day-to-day is the same for all.

This assumption can be and will be checked against the data later (it passes). The remaining assumptions are that the skills of the players entered—the  $S_i$ —are distributed independently as a normal distribution with mean 0 (the daily par having been subtracted) and variance  $\sigma_S^2$ , and that the  $L_{ij}$  are distributed independently as a normal distribution with mean 0 and variance  $\sigma_L^2$ . The daily pars  $\mu_j$  are considered to be four constants, different in each tournament.

The analysis is based upon the recorded player's scores in a large number of tournaments. We estimate, separately for each tournament, the  $\mu_j$  and the variances  $\sigma_S^2$  and  $\sigma_L^2$ . Note that  $\sigma_S^2$  is the variance of skills of all entering players, regardless of whether they eventually make the cut.

The interest in this model comes from the ease of interpreting the results from these estimates. For one player in one round, we have:

$$X_{ij} = \mu_j + S_i + L_{ij}.$$

This, under the model's assumptions of independence, gives a simple expression for the relative contributions of the two sources of variability for a randomly selected player:

$$\begin{aligned} \text{Variance}(X_{ij}) &= \text{Variance}(S_i) + \text{Variance}(L_{ij}) \\ &= \sigma_S^2 + \sigma_L^2. \end{aligned}$$

The two variances may differ, but they are weighted equally. However, over a four-day tournament, the luck “averages out” to a degree and the player's variability in total score— $\sum_j X_{ij}$ —has a different breakdown:

$$\sum_j X_{ij} = \sum_j \mu_j + 4 S_i + \sum_j L_{ij}.$$

and

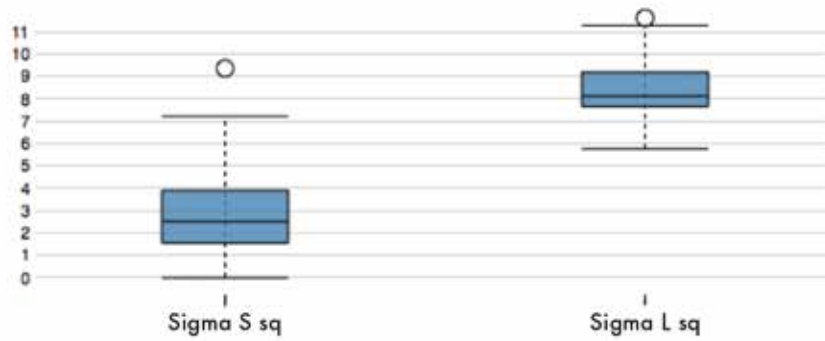
$$\begin{aligned} \text{Variance}(\sum_j X_{ij}) &= \text{Variance}(4 S_i) + \text{Variance}(\sum_j L_{ij}) \\ &= 16\sigma_S^2 + 4\sigma_L^2. \end{aligned}$$

Here, the weight of skill over luck is 4 to 1. We will discuss the implications of these relationships in the light of data collected over more than 20 years.

## The Data

The primary data are the results of all four “major” men's tournaments—the Masters, PGA, and U.S. and British Opens—for all 22 years from 1994 through 2015. In addition, we looked at data for a less-complete set of major women's tournaments: the





	<b>Skill Var</b>	<b>Luck Var</b>
Ave BO	1.51	8.54
Ave Masters	3.02	8.50
Ave PGA	3.38	8.16
Ave USO	2.13	8.73
Ave Men	2.51	8.48
Ave WPGA	3.85	6.99
Ave WUSO	4.39	8.67
Ave Women	4.12	7.83
Ave All	2.81	8.36
Min	0.00	5.75
LQuartile	1.56	7.65
<b>Median</b>	<b>2.50</b>	<b>8.13</b>
UQuartile	3.91	9.18
Max	9.35	11.61
IQRRange	2.35	1.53
Range	9.35	5.85

Figure 2. Estimated skill and luck variances for the 108 tournaments, summarized.

Women's PGA for 10 years (2001–2009 and 2014) and the Women's Open championships for 10 years (2001–2009 and 2013).

In every case, all golfers who completed the tournament were included—"completed" meaning they were either cut after two full rounds or finished all four rounds. Thus, in the 2005 U.S. Open, 156 golfers completed round 1, but only 154 completed the tournament, the other two dropping after round 1 and before the cut. In every tournament, the daily

pars and skill and luck variances were estimated by maximum likelihood.

The estimated skill and luck variances for the 108 tournaments are summarized and displayed in Figure 2.

The first striking feature is the remarkable lack of variation in the luck variance: In a full 50% of the cases, the estimates fall between 7.65 and 9.18; the median is 8.13 and the mean is 8.36. The luck variance is essentially the same for men, women,

**Table 1—Luck vs. Skill**

	<b>Luck Exceeds</b>	<b>Skill Exceeds</b>
BO	15	7
Masters	9	13
PGA	3	19
USO	14	8
WPGA	0	10
WUSO	1	9

and in all tournaments. As shown later, it is also the same from the top golfers down through the bottom quartile.

The second striking feature is how large this variance is: A median variance of 8.13 gives a standard deviation of 2.85, which suggests a luck variation of plus or minus 5.5 strokes a round is not unusual. The irreducible luck component is large and may be expected to have a considerable effect upon outcomes.

What about skill? The situation there is mixed in interesting ways. The smallest skill variance is in the British Open; the largest in the women's tournaments (both individually and in aggregate). In fact, of the 22 British Opens, two have skill variance estimated to be zero, and three others are nearly as small. We suggest two explanations for this.

The first is that the British Open is typically held on narrow courses bounded by treacherous rough and with numerous hazards that are not familiar on the PGA tour in North America. If we add to that the frequently unfriendly character of the weather on the Scottish coast, the skills perfected on the calmer, manicured courses in America may just be less relevant. The course may not be physically level, but the "playing field" of relevant skills may be unusually so.

The other explanation may be related to weather also: The model we use makes no allowance for changes in conditions during a round that may favor early or late starters. Changes between days are allowed by the model through the par, but not within days, and to the degree this happens, the model would attribute a good part of the change to luck.

We suspect both causes are present there in different years. Indeed, the two Opens with skill variance estimated to be zero (2008 and 2013) have the two largest estimated luck variances in any of the men's tournaments: 11.05 and 11.27. The model may simply fit the British Open less well.

The larger skill variance for the women's tournaments seems likely to be tied to differences between

the two tours. The women's tour is much smaller than the men's, presumably due to its being much more poorly compensated, and so it is far more difficult for women to reach a subsistence level on tour. The 2017 purse for the women's U.S. Open was \$5 million, the largest purse of any women's tournament (the WPGA the same year was \$3.5 million), but small compared to the Men's U.S. Open that year at \$12 million.

While all the winners of the women's tournaments do well (top prizes of \$500,000 or so; still much less than the men, who may take over \$2 million), the lowest prize among those who made the cut for the 2017 Women's U.S. Open was less than \$7,000, not even close to covering expenses, and those who missed the cut got nothing.

The top players on both tours are superb golfers who can continue to survive, but the middle and lower players on the women's tour can only remain for a short time unless they meet success. The men have many more tournaments, and richer prizes, and while it is also difficult to survive in that tour, many more enter and only the best of these make it to the majors. The women's majors draw from a smaller pool, and must draw more deeply.

As mentioned earlier, the variance of total scores over four rounds is  $16\sigma_S^2 + 4\sigma_L^2$ , which weights skill over luck by 4 to 1: Skill will outweigh luck if 4 times the skill variance exceeds the luck variance (see Table 1).

Bear in mind that "skill" here means range of skills among *initial* entrants, not absolute skill, and that certain types of within-day weather changes (for example, as in some British Opens) that favor early starters will contribute to variation as "luck" for our model.

By running over four days, a golf tournament balances the contributions of skill and luck. Table 2 shows how this goes, "ratio" being  $= (\# \text{rounds}) \times \sigma_S^2 / \sigma_L^2$ .

For men's tournaments, a one-day/one-round plan would have luck variation exceeding skill variation by

**Table 2—Contributions of Skill and Luck**

	<b>Skill Variance</b>	<b>Luck Variance</b>	<b>Ratio (1rd)</b>	<b>Ratio (2rd)</b>	<b>Ratio (3rd)</b>	<b>Ratio (4rd)</b>
Ave BO	1.51	8.54	0.18	0.35	0.53	0.71
Ave M	3.02	8.50	0.36	0.71	1.07	1.42
Ave PGA	3.38	8.16	0.41	0.83	1.24	1.66
Ave USO	2.13	8.73	0.24	0.49	0.73	0.98
<b>Ave Men</b>	<b>2.51</b>	<b>8.48</b>	<b>0.30</b>	<b>0.59</b>	<b>0.89</b>	<b>1.18</b>
Ave WPGA	3.85	6.99	0.55	1.10	1.65	2.20
Ave WUSO	4.39	8.67	0.51	1.01	1.52	2.02
<b>Ave Women</b>	<b>4.12</b>	<b>7.83</b>	<b>0.53</b>	<b>1.05</b>	<b>1.58</b>	<b>2.10</b>

**66 Players' Luck SD vs. Order of Finish**

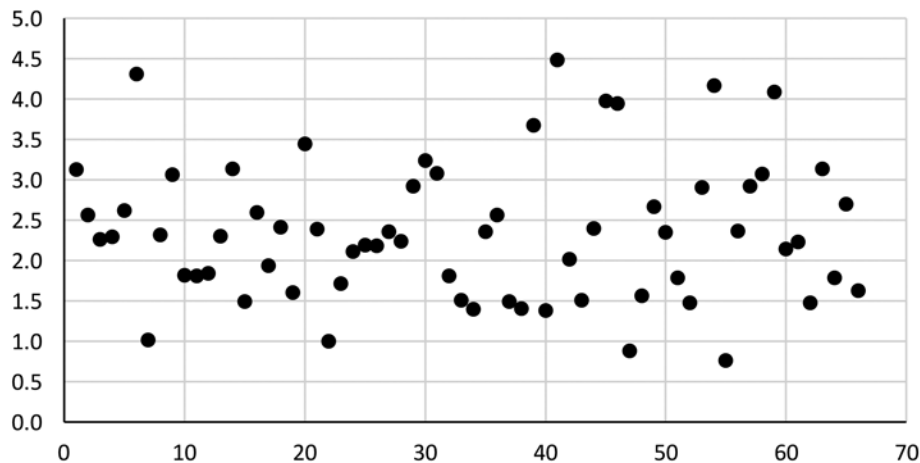


Figure 3. The vertical axis gives the estimated standard deviation per 18-hole round for each player; the horizontal axis gives the order of finish by total score for the 66 players who competed in both the 2005 Masters and PGA; the leftmost point is Tiger Woods. There is clearly no marked trend.

more than 3 to 1; four days of play moves the balance considerably to slightly on the other side. Clearly, the balance varies: It is much more weighted toward luck for the British Open, as noted earlier (where the effect of weather plays a role, too), and much more toward skill in women's tournaments. However, since the 1930s, most tournaments have run for 72 holes played over four days. It seems reasonable to speculate that 80 years ago, the skill variation for the men would have been more like that for women today: The number of professional men then was much smaller and the purses much lower.

The need to balance skill and luck requires a tournament to run several days; for the better part of a century, the four-day, 72-hole tournament has served as a convenient compromise.

## The Constancy of Luck Variation

Our simple model supposes that luck variation is the same for all players, from the best to those who fail to make the cut. To see if the data support this assumption, we looked at all 66 golfers who entered both the Masters and the PGA in 2005. We ordered them according to their total scores for the two tournaments. Players 1 to 32 made the cut in both tournaments, and players 59 to 66 missed the cut in both; the others (33 to 58) made the cut in one but not the other.

Player #1 was Tiger Woods; he won the Masters and finished third in the PGA that year. Player #2 was Phil Mickelson; he won the PGA and finished ninth in the Masters. For each player, we estimated the standard deviation per round (after subtracting

## TECHNICAL APPENDIX

The model employed in the analysis of each tournament is a mixed random effects model with missing data: The scores for the last two rounds of those players who miss the cut are necessarily (by design) not available. By a linguistic quirk, they satisfy the technical definition for “missing completely at random,” since the missingness is determined entirely by the data that are observed, and the model permits, with all available data, writing down and directly maximizing the likelihood, without needing to compensate for possible data-dependent mechanisms leading to missing data.

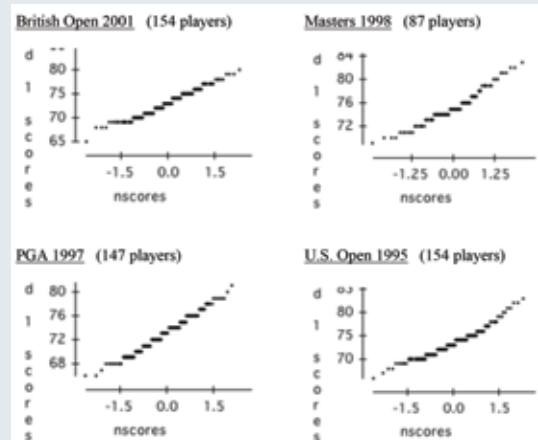
The fitting uses the EM Algorithm, where the true skills  $S_i$  are also considered as missing data (Little and Rubin, 1987, pp. 129, 149–151). The iteration converged rapidly except where the skill variance was at or near the boundary value of 0; in those cases, we refit with the boundary value to ensure we had found the maximum.

In all cases, we estimated the estimate’s variances and covariances. We also fit but did not report data for a number of non-major men’s tournaments and the results were much the same as for the majors, especially regarding the constancy of luck variance.

Of course, the model is simplistic; we noted some lack of fit in cases where there were anomalous weather patterns or rain delays, but except for the effect noted (particularly in a few British Opens) where weather augmented the “luck” variation, we were generally satisfied with the fit.

Golf scores are necessarily integer-valued and cannot be strictly normally distributed. Mosteller and Youtz have modeled them as Poisson on a shifted base (1992, 1993), but that too was an approximation, and our data pass the scrutiny of several diagnostic tests for normality (for example, normal probability plots show approximate linearity and a lack of noted skewness, as would be expected with normally distributed data).

Here are some examples of normal probability plots for the first rounds of typical years for four tournaments.



As can be seen, the first round scores for each tournament in the selected years are in straight lines for the most part.

For Greg Norman’s regression equation, we have, in terms of our model,  $(X, Y)$  as bivariate normal with means  $\mu_X = \mu_1 + \mu_2$  and  $\mu_Y = \mu_3 + \mu_4$ , equal variances  $\text{Var}(X) = \text{Var}(Y) = 4\sigma_s^2 + 2\sigma_L^2$ , and covariance  $\text{Cov}(X, Y) = 4\sigma_s^2$ , giving the correlation  $\rho_{XY} = 2\sigma_s^2 / (2\sigma_s^2 + \sigma_L^2)$ . Given the total score  $X$ , the conditional expectation for  $Y$  is then  $E(Y|X = x) = \mu_Y + \rho_{XY}(x - \mu_X)$ . The data for the 1996 Masters give the maximum likelihood estimates of  $\mu_1, \mu_2, \mu_3, \mu_4$  as 73.4022, 74.4130, 74.9044, 74.8362, and the maximum likelihood estimates of  $\sigma_s^2$  and  $\sigma_L^2$  as 3.8794 and 9.1377, so  $\rho_{XY}$  is  $2 \times 3.8794 / (2 \times 3.8794 + 9.1377) = 0.4592$ . Now  $x = 63 + 69 = 132$ , so  $E(Y|X = 132) = 74.9044 + 74.8362 + 0.4592(132 - 73.4022 - 74.4130) = 142.4782$ .

The small simulation referred to in the section on “The Constancy of Luck Variation” was done in Excel. Recall that according to our simple model, the total score of a player  $i$  for four rounds is:

$$T_i = \sum_j X_{ij} = \sum_j \mu_j + 4 S_i + \sum_j L_{ij}.$$



## TECHNICAL APPENDIX (CONTINUED)

We created a simulated tournament with 150 entries, first with an ordered column of  $280 + 4S_i$  (taking  $\sum_i \mu_i = 280$  as if par = 70 each round) using an approximation to the expected values of Normal random variables with means 280, and variance  $16\sigma_S^2 = 16 \cdot 2.5$  (16 times the median skill variance from our data). We then added a luck column that was random Normal with mean 0 and variance 4 times the median luck variance for a single round (since  $\text{Var}(\sum_i L_{ij}) = 4\sigma_L^2 = 4 \cdot 8.13$ ).

If the two columns were added, they would give the results for a simulated four-round tournament (where the “cut” is ignored). This was repeated 250 times and the finishing ranks of the 150 participants computed for each simulated tournament. For the purposes of this study, the highest skill was changed to be the second skill minus 1 and the tournament repeated 250 times; then similarly but with minus 2. This provides the results for both 1- and 2- stroke advantages.

the estimated daily par for each completed round, we estimated the variance for each player for each tournament, pooled these with appropriate weights, and took the square roots). The display shows no pronounced violation of the assumption; the increased scatter for the lower-ranked players reflects the fact that fewer data were available from tournaments where the cut was missed.

A natural question to ask is how likely is the golfer with the highest level of skill to win a tournament? That turns out to be not quite the right question: The chance the most-skillful player will win depends tremendously upon the gap between the top player and the second-most-skillful player, and upon the tightness of the grouping of skills among the others behind the second.

Instead, we looked at the dependence of the probability of winning on the gap between #1 and #2, supposing the players below #2 had an average configuration for a Men’s PGA tournament (see Technical Appendix). In a small simulation study using the median values for our data of  $\sigma_S^2 = 2.50$  and  $\sigma_L^2 = 8.13$ , we found that if the top player has an average 1-stroke advantage over #2 (i.e.  $S_1 = S_2 - 1$ ), the probability of winning is about 31%; the chance of finishing second is about 16%. If the advantage is increased to an average 2-strokes ( $S_1 = S_2 - 2$ ), the chance of winning rises to about 58% and the chance of finishing second is about 16%. To get a feeling of what this means, a one-stroke advantage is quite large at this level of play. Under our model, the expected gap between the top two entrants is only 0.4 strokes, and the top entrant is expected to win 16% of the time.

A two-stroke advantage is almost unheard of. It is said that at his very best, Tiger Woods played an average of about two strokes a round better than the nearest competitor, a level he could not sustain for long.

### Regression

This study began in April 1996, initially to answer a simple question. That was the year golfer Greg Norman went into the final round of the Masters with a six-shot lead, but he lost the tournament, ending up five strokes behind Nick Faldo. In the process, he became a symbol of “choking.” The next day’s *Chicago Tribune* headline stated “Norman Gags.” Indeed, his descent was steady from the beginning; the top of the leaderboard after the fourth round looked like this:

N. Faldo	69-67-73-67 – 278	–12
G. Norman	63-69-71-78 – 281	–7
P. Mickelson	65-73-72-72 – 282	–6

If you were to plot Norman’s four rounds, they show an almost-linear trend for the worse. But was Norman really unusual in his performance? Many others in the same tournament showed a similar (but less-severe) trend. In fact, of the 44 who finished, 33 showed a net trend for the worse and 9 showed a trend for the better. Two golfers showed no trend at all, including the eventual winner, Nick Faldo.

The course did play a bit harder in the later rounds, but correcting for that has little effect; 30 of the 44 still trended for the worse.

We suspected that much of this was simply evidence of the “regression to the mean” phenomenon: If two variables  $X$  (a golfer’s score on the first two rounds) and  $Y$  (same golfer for the last two rounds) are imperfectly correlated, and an individual is selected on the basis of an extreme value of  $X$ , then we should, on average, expect that in standard deviation units,  $Y$  should be closer to the population average than is  $X$ .

The 44 who finished the 1996 Masters were selected as (roughly) the top half of the entering field of 92, based upon their total score in the first two rounds. We should expect an apparent drop in performance for the second half. This is true even supposing (as we do) that their skill remains intact: If their first-half performance was due in part to good luck, that component could not be expected to continue; they would go from “Good Skill + Good Luck” to “Good Skill + Average Luck.”

The regression phenomenon surely played a role for the field as a whole, but was it enough to account for Norman’s fall from grace? In terms of our simple model,  $(X, Y)$  are bivariate normal with means  $\mu_X = \mu_1 + \mu_2$  and  $\mu_Y = \mu_3 + \mu_4$ , variances  $\text{Var}(X) = \text{Var}(Y) = 4\sigma_S^2 + 2\sigma_L^2$ , and correlation  $\rho_{XY} = 2\sigma_S^2/(2\sigma_S^2 + \sigma_L^2)$ . Given the total score for the first two rounds and assuming “no choking” (so the skill remains the same), the conditional expectation for the last two rounds is then  $E(Y|X = x) = \mu_Y + \rho_{XY}(x - \mu_X)$ . Using the data for the 1996 Masters to give estimates of the means and correlation, we have that for Greg Norman,  $x = 63 + 69 = 132$ , so  $E(Y|X = 132) = 149.7 + 0.459(132 - 147.8) = 142.45$ , but his actual score for the last two rounds was  $71 + 78 = 149$ , or 6.5 strokes worse than the regression predicted.

Norman lost by 5 strokes; had he only regressed as expected, he would have won. On the other hand, 6.5 strokes is only 1.5 “luck” standard deviations, so the evidence for choking, while quite suggestive, is not absolutely compelling.

## An Interesting Contrast

Tournament golf and several other sports balance the effects of skill and luck by adjusting the length of the tournament. NCAA basketball also structures its major tournament (“March Madness”) for such a balance, but they do it in a totally different manner. The basic tournament (after some minor stages to allow a few marginal teams to enter late) involves 64 teams that are thought by the organizers to be the best in the USA. This choice is made very late in the season, when there is a great deal of information on their ability; it is not a random selection.

In a way, their problem is that there is too much information on skill: If 10 experts were to independently each name for whom they thought were the 10 best teams, there would be a considerable overlap. Why is this a problem? Because there are too many teams to allow the same two teams to play multiple times, and in any single game, if two teams play that are at all similar in skill, there is a nontrivial chance the weaker will beat the stronger.

If any two teams in the top half of the 64 were to play a single game, the data suggest the chance of an upset is above 15%. For the top quarter, it is above 30%. The organizers evidently worried that if the structure used random pairings (so, for example, the top two teams might play each other in the first round), there would be a good chance that the best teams would be knocked off early, and fan interest and the credibility of the tournament would suffer. The effect of luck was too strong.

To deal with this, the NCAA divided the field of 64 into four evenly matched divisions of 16 teams each, and in each division, they ranked (“seeded”) the teams in the organizers’ judgment from 1 (best skill) to 16 (lowest skill). They then set up a schedule of play (“the bracket”) that would determine the paired match-ups; each division would produce a winner; these would then be paired for the semifinal matches; and the one final match between the semifinal winners would decide the championship. In the first round in each division, the #1 seed would play #16 seed, #2 seed plays #15, ..., and #8 seed plays #9 seed. In the second round, the winner of #1 vs. #16 plays the winner of #8 vs. #9, and so forth.

The goal was to try to limit the chance a top-seeded team would be eliminated early, and this part has been fairly successful: Through 2017, no #16 has ever beaten a #1 since this version of the structure was introduced in 1985 (in 2018, the overall favorite was upset, to everyone’s surprise). On the other hand, the #8 vs. #9 game is even (53% wins for #9). The second round has gone to the #1 seeded team in the division about 85% of the time.

If the plan was to reduce the role of luck in the early rounds of the tournament, it was moderately successful. Over the 26 years from 1985–2010, 80% of those who survived the third round were initially seeded among the top four in their division. With random assignment of opponents, we would expect only about 70% of those survivors to initially be among the top four, using past data to estimate the chances of one seed beating another. But the final rounds are another matter altogether: There, the opponents

are much more evenly matched (the flip side of what the structure led to in the early rounds), and the role of luck then becomes magnified.

In March Madness, the role of luck was, by design, shifted to the late rounds—and that of skill to the early rounds; a very different balance from golf. To sustain fan interest, the top seeds have a protected status early on, and the result is as predicted: In the 26 years from 1985–2010, the champion was one of the four #1 seeded teams 16 times. The prize went to one of 12 teams seeded among the top three in one of the four divisions 23 times out of 26.

## Conclusions and Speculations

It has not escaped our notice that the balance we find between skill and luck in tournament golf may be expected in other sports as well, and in many social practices beyond sports. No sport where luck dominates can be expected to hold the public's interest: Lotteries are all luck and attract no spectators and, even though the sums involved can cause a stir, there is no consequential interest in who wins beyond the winner's immediate family and predators eager to help the winner invest.

At the other extreme, no sport where skill is paramount will consistently attract much interest. Chess comes closest, but even there, it is the potential for human error that holds the occasional small crowd. A contest between two computers will not even be watched by other computers. The twin lures of human excellence and uncertainty are what rule. Just as the structure of a golf tournament is presumably intended to generate interest (and consequently income for all involved), so too in baseball, basketball, tennis. In all these situations, the questions would be what is the structure of tournaments and what considerations lie behind what might seem like arbitrary choices?

Similar situations arise in the arts and sciences. Consider the Oscars and the Nobel Prizes. In each

case, an award goes to a few from among many highly skilled individuals eligible. In most cases, there is uncertainty beforehand about who will win. The interest in the year's awards would be minimal if the choices were either predictable with certainty (skill dominates) or made randomly and completely unpredictably (luck dominates). In these examples, the tournament structure is reasonable opaque, leading occasionally to accusations of bias, but the effect—the mixture of skill and luck—is much the same as that of a structured sports tournament. ■

## Further Reading

- Mosteller, F., and Youtz, C. 1992. Professional Golf Scores are Poisson on the Final Tournament days. *1992 Proceedings of the Section on Statistics in Sports*, 39–51. Washington, DC: American Statistical Association.
- Mosteller, F., and Youtz, C. 1993. Where Eagles Fly. *CHANCE* 6:37–42.
- Little, R.J.A. and Rubin, D.B. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

## About the Authors

**Margaret Stigler** has a master's degree in sports administration from Northwestern University, with a thesis, "Statistics at the Gate," that examined the attendance trends for baseball parks across the U.S. She now works on data analytics at the Center for Research Libraries in Chicago. She is a more skillful golfer than her father Stephen, who relies more on luck on the golf course.

**Stephen Stigler** is professor of statistics at the University of Chicago. His latest book is *The Seven Pillars of Statistical Wisdom* (Harvard University Press, 2016). This study is based on collaboration over more than a decade not funded by any government or other agency.



# Why a 1-for-45 Record in the Kentucky Derby Does Not Necessarily Equate to Underachievement

*Leonard Cupingood*

As of May 1, 2017, at age 49, Todd Pletcher was, by one measure (total purse earnings), the most-successful trainer in thoroughbred horse racing history, with over \$300 million in earnings. By another measure (career wins), Pletcher ranks 10th. He also received the Eclipse Award for Outstanding Trainer seven times from 2004 to 2014 (Eclipse

awards are determined in annual voting by the National Thoroughbred Racing Association, *Daily Racing Form*, and National Turf Writers Association). However, before the running of the 2017 Kentucky Derby, the media were widely reporting his prior record in that event with the implication that it was notably subpar. Dick Jerardi, a veteran horse-racing

journalist, wrote that “Pletcher is 1-for-45 in the Kentucky Derby. It is one of the more bizarre statistics in the sport.”

This 1-for-45 statistic was referring to the fact that Pletcher was the trainer of 45 horses in 16 previous runnings of the Kentucky Derby (someone can be the trainer of more than one horse in the same race). Even Pletcher



himself was quite aware of this seemingly adverse statistic. In the *New York Times* coverage of the 2017 Kentucky Derby, Joe Drape wrote that “Pletcher was 1-for-45 in the Derby when the gates popped open, a number that even he found troubling.”

Pletcher was the trainer of an average of  $45/16 = 2.82$  horses in 16 previous Kentucky Derbies. Since there is only one winner of each race, Pletcher could only have won 16 of them, not 45, so before 2017, it was more realistic to say he was really 1-for-16. Moreover, since all horses are not equally likely to win any given race, even a 1-for-16 statistic is misleading. How should we evaluate whether this “1-for-45” performance is better, worse, or just about what would be expected?

This article uses estimated probabilities of winning each race derived from the wagering to calculate how many Kentucky Derbies Todd Pletcher would statistically have been expected to win. Then, using these same estimated probabilities of winning each race, I estimate a distribution of his number of previous Kentucky Derby wins through simulations of the winning horse for all Kentucky Derbies in which Pletcher participated as a trainer. I then update the statistical expectation and distribution of the number of Pletcher wins to incorporate the outcome of the 2017 Kentucky Derby.

### Estimated Probabilities for Each Horse to Win a Race

As noted above and should be self-evident, every horse in a race is not equally likely to win. Studies have shown that, for the most part, the relative percentage of dollars bet on each horse closely measures each horse’s individual chance of

winning the race. An exception to this is the favorite-longshot bias, which refers to the situation where bettors tend to over-bet on longshots and under-bet on favorites (see Hausch, Lo, and Ziemba).

There are many different possible wagers in a race, each with a separate wagering pool, e.g., win pool for winning the race; place pool for finishing first or second; show pool for finishing first, second, or third; and other exotic wagers, such as the exacta and trifecta, which involve the exact finishing position for multiple horses in the same race. This article focuses only on the winner of the race.

Odds and probabilities are terms that are closely related. If the odds against an event occurring are 2-to-1, then on average, it will *not* occur twice as often as it will occur, and its probability of occurrence is  $1/3$ . In general, the probability of occurrence is  $1/(1 + \text{odds})$ . In horse racing, the relative chances of each horse winning a race are traditionally quoted in terms of odds, not probabilities. However, these odds reflect “payout odds,” which determine how many dollars are won per dollar wagered on the winning horse. These “payout odds” differ from the estimated “true odds” of a horse winning a race due to the track first deducting a percentage of the dollars wagered before paying the remainder to those with winning wagers. I call this deduction a track retention percentage, and denote it as  $R$ .

It helps to illustrate how the final win payout odds for each horse are determined, and then how each horse’s probability of winning can be estimated from these payout odds. This conversion from payout odds to estimated win probability is important because the payout odds for a horse are available on a historical basis, but the information about the amount

of dollars bet on a horse is not readily available.

Assume there are three horses in a race with respective final wagered amounts of money in the win pool of \$5,000 on Horse A, \$3,000 on Horse B, and \$2,000 on Horse C. The probabilities of each horse’s chances of winning are estimated by the relative proportion of dollars bet on each horse to win the race. In this example, these probabilities are 0.50 for Horse A, 0.30 for Horse B, and 0.20 for Horse C. For win betting, the final payout odds for a given horse are determined by a) reducing the total dollars wagered in the win pool by the percentage  $R$ , b) then subtracting the dollars bet on each given horse from this reduced amount, and c) dividing this result by the dollars wagered on each given horse.

In the example above, bettors wager a total of \$10,000 on Horses A, B, and C to win. Using an assumed track retention percentage  $R$  of 20%, \$8,000 is available to be paid to those who bet on the winning horse. The payout odds for each horse in this example are:

Horse	Dollars Wagered	Payout Odds
A	\$5,000	$(\$8,000 - \$5,000) / \$5,000 = 0.60$
B	\$3,000	$(\$8,000 - \$3,000) / \$3,000 = 1.667$
C	\$2,000	$(\$8,000 - \$2,000) / \$2,000 = 3.00$

In this example, 50% of the dollars wagered are on Horse A. Assuming the relative percentage of dollars wagered is the determinant of true odds, Horse A should have true odds of winning the race of 1-to-1, a probability of 0.50. If

the track did not retain any of the dollars wagered ( $R = 0$ ), the payout odds would in fact be  $(\$10,000 - \$5,000) / \$5,000 = 1.00$ , or 1-to-1. It can be shown that a horse's estimated win probability based on the wagering can be obtained from the payout odds by  $(1 - R) / (1 + \text{odds})$ . The estimated win probability of 0.50 for Horse A, using a track retention percentage  $R$  of 20%, is obtained as  $(1 - 0.20) / (1 + 0.60) = 0.50$ .

As noted above, while the actual dollar amounts bet on each horse are not readily available on a historical basis years after a race has been run, the final payout odds for each horse are. I estimated the win probabilities for the previous Kentucky Derbies using the published final payout odds from the historical charts, and the track retention percentages in effect for each race.

## Entries and Fields

Before 2001, two situations in the Kentucky Derby provide a slight complication for assigning win probabilities to each trainer. These are situations where more than one horse is "coupled" for betting purposes, which means that the odds on the "coupled" horses pertain to the entire group of horses that were coupled.

The first situation, "entries," occurs when more than one horse has the same trainer or owner. If the focus of the analysis is on the trainer's chances of having the winning horse, it is straightforward to assign the odds of the coupled entry to the trainer when each coupled entry comprises horses with the same trainer. There were 12 situations from 1991 to 2000 with coupled entries, and 11 of these 12 were due to horses having a common trainer (there can be more than one coupled entry in a race).

The second situation arose when the total number of

separate betting entities was limited to a fixed number, typically 12. If there were 15 horses in a race limited to 12 betting interests, then 11 individual horses represent 11 separate betting interests and the "12th betting interest" comprised the other four horses as a group, known as the "field." If these four field horses are trained by four different trainers, it is not clear how to partition the group odds for the field entry into separate odds for the four individual trainers. Since the horses grouped into the field were typically those judged to have the least chance to win, it is not unreasonable to divide the collective field win probability into equal parts, and assign these equal fractional portions of the win probability to each trainer of a horse in the field. I followed this operational rule for assigning fractional win probabilities to trainers having field horses.

Beginning in 2001, the number of betting interests increased up to a maximum of 20, with individual betting interests for each horse in the Kentucky Derby, regardless of whether an owner or trainer had an interest in more than one horse in the race.

## Analysis of Expected Number of Trainer Wins

The first Kentucky Derby ran in 1875 and it was repeated every year thereafter, making 2017 the 143rd running of the race. When studying trainers in a particular race that only runs once a year, the number of observations for a particular trainer is necessarily limited by the number of years in that person's career of training horses.

Also, while a trainer may be fortunate enough to have more than one entry in the Kentucky Derby in a particular year, the trainer will have no entries

in other years. I extracted all available Kentucky Derby chart data from [www.equinbase.com](http://www.equinbase.com), which covered the years 1991–2017. These charts include the order of finish, trainer and jockey for each horse, final payout odds, and coupling and field entry information.

During the 26-year time period excluding 2017, only five trainers had a total of 10 or more entries in the Kentucky Derby: Pletcher (45), D. Lukas (30), Bob Baffert (27), Nicholas Zito (25), and Steve Asmussen (15).

Table 1 presents the frequency distribution of the number of horses entered by a trainer in the Kentucky Derby from 1991 to 2016.

Collectively, 198 different trainers entered horses in the Kentucky Derby from 1991 to 2016. For more than 2/3 of these trainers (68.2%), a single Kentucky Derby entry represented the only entry over these 26 years. Almost 95% of the trainers (186 out of 198) had five or fewer entries during these 26 years.

To analyze how well each trainer performed statistically, I calculated the probability for each Kentucky Derby that a trainer would win the race by summing the probabilities of all the trainer's horses in the race. The probabilities were computed from the final payout odds of each horse in the race for a single given year. The expected number of wins for the trainer in a given year is simply  $nP_t = P_t$ , where  $P_t$  is the trainer's aggregate probability of training the winning horse in the race and  $n = 1$  for a single year.

The expected number of total wins for the trainer is obtained by summing the trainer's individual year probabilities over all 26 years, since the outcomes of winning in one year should be statistically independent of winning in any other year. I then computed the surplus in wins for each trainer as

**Table 1—Frequency Distribution of Number of Entries by Trainer in the Kentucky Derby, 1991–2016**

Number of Trainer Entries	Frequency	Percent
45	1	0.5%
30	1	0.5%
27	1	0.5%
25	1	0.5%
15	1	0.5%
9	1	0.5%
7	3	1.5%
6	3	1.5%
5	8	4.0%
4	3	1.5%
3	17	8.6%
2	23	11.6%
1	135	68.2%
<b>Total</b>	<b>198</b>	<b>100.0%</b>

*In more than 2/3 of the cases from 1991–2016, a trainer entered only a single horse, while five trainers had 10 or more entries each.*

**Table 2—Trainers with Expected Number of Wins Exceeding 1 in Absolute Value, 1991–2016**

Trainer	Number of Races	Entries	Expected	Wins Actual	Surplus
Doug O'Neill	4	5	0.45	2	1.55
D. Lukas	18	30	1.47	3	1.53
Bob Baffert	17	27	2.67	4	1.33
Todd Pletcher	16	45	2.16	1	-1.16

*From 1991–2016, only four of 198 trainers had a surplus in expected number of wins exceeding 1 in absolute value.*

the actual number of wins minus the expected number of total wins. Clearly, the total surplus for virtually all of the trainers will be small fractions, since the vast majority of trainers, as noted above, have trained horses in five or fewer Kentucky Derbies. Only four trainers

had a surplus that exceeded 1 in absolute value for the 26-year time period studied.

Pletcher is the only trainer with a shortfall in expected number of wins exceeding 1.

However, the total number of expected wins for his 45 entries in

16 races is only 2.16. This reflects the fact that in most years, his horses were largely longshots: horses with low probabilities of winning. In only three of the 16 years did his probability of winning a given Kentucky Derby exceed 0.20.

**Table 3—Distribution of the Number of Wins by Todd Pletcher in the Kentucky Derby 2000–2016, Based on 1,000 Simulations**

Number of Wins	Frequency	Percent
0	78	7.8%
1	263	26.3%
2	303	30.3%
3	202	20.2%
4	101	10.1%
5	37	3.7%
6	13	1.3%
7	2	0.2%
8	1	0.1%
<b>Total</b>	<b>1,000</b>	<b>100.0%</b>

*Todd Pletcher's actual win total of 1 win for the 2000–2016 time period is not a rare event.*

**Table 4—Distribution of the Number of Wins by Todd Pletcher, Including the 2017 Kentucky Derby, Based on 1,000 Simulations**

Number of Wins	Frequency	Percent
0	56	5.6%
1	228	22.8%
2	296	29.6%
3	218	21.8%
4	124	12.4%
5	54	5.4%
6	17	1.7%
7	6	0.6%
8	1	0.1%
<b>Total</b>	<b>1,000</b>	<b>100.0%</b>

*After including the result of the 2017 Kentucky Derby, Pletcher's two wins are the most likely expected occurrence.*

How should we evaluate Pletcher's record in these 16 Kentucky Derbies? To help judge how likely it is that Pletcher would win exactly one Kentucky Derby with these 45 entries in 16 races, I conducted 1,000 simulations of

all the Kentucky Derbies from 2000 to 2016 in which Pletcher was a trainer of at least one horse in the race. The winner of each simulated race was based on each trainer's probabilities of winning the race. Table 3 gives the resulting

distribution of the number of Pletcher wins.

The actual outcome—one Pletcher win—is not an unlikely event, occurring 26.3% of the time. Even winning none of these Kentucky Derbies would also not have



been that unusual, an event that might be expected to occur about 8% of the time. This means that the one-tail probability of Pletcher's "1 or fewer wins" for his 16 Kentucky Derbies might be expected to occur approximately 34% of the time. This is roughly comparable to the one-tail probability associated with flipping a fair coin 50 times and observing 23 heads—two fewer than expected. No one should reasonably complain about the fairness of a coin that "only" produces 23 heads in 50 flips.

## Updating Todd Pletcher's Record After the 2017 Kentucky Derby

In 2017, Always Dreaming, a Todd Pletcher-trained horse, won the Kentucky Derby. Pletcher was the trainer of two other horses in the 2017 Kentucky Derby, bringing his record to two wins in 17 Kentucky Derbies in which he trained a total of 48 horses. Re-calculating the expected number of wins including 2017, Pletcher's shortfall in total wins drops to 0.39. Table 4 incorporates the results of an additional 1,000 simulations of the

2017 Kentucky Derby, and presents an updated distribution of the total number of Pletcher Kentucky Derby wins.

Now, the most likely number of wins is two—exactly the number of wins achieved by Pletcher.

## Conclusion

The analysis provides an illustration of how a widely quoted statistic used to imply under-achievement can be misleading, and not at all support the allegation for which it was used. Here, the reference to a 1-for-45 record in the Kentucky Derby clearly implies that an otherwise highly successful trainer has under-achieved in his attempts to win the race.

However, when statistically analyzed with respect to the often-low probabilities of Pletcher's entries in the Kentucky Derby, we see that this 1-for-45 statistic actually only represents an expected shortfall of 1.16 wins, an event that is not particularly rare. Now, as the winning trainer of the 2017 Kentucky Derby winner, Pletcher's win total of two is statistically the most likely observed win total. ■

## Further Reading

Drape, Joe. At Kentucky Derby, Always Dreaming Produces a Winner's Circle Reunion Party. *www.nytimes.com/2017/05/06/sports/kentucky-derby-always-dreaming.html*.

Jerardi, Dick. Always Dreaming of a Derby Win, Philly.Com, May 6, 2017. *www.philly.com/philly/columnists/dick\_jerardi/Kentucky-Derby-Always-Dreamingtrainer-Todd-Pletcher.html*.

Rausch, Donald B., SY Lo, Victor, and Ziemba, William T. (eds.). 2008. Efficiency of Racetrack Betting Markets, Part IV. Efficiency of Win Markets and the FavoriteLongshot Bias (pp. 251–352). *World Scientific*.

## About the Author

**Leonard Cupingood** is a director at BLDS, LLC, a consulting firm primarily specializing in statistical analyses for litigation support. He has a PhD in statistics from Temple University and is also a long-time horseracing enthusiast and fractional owner of thoroughbred horses with West Point Thoroughbreds, Inc.



# All in the Family: German Twins' Finishing Times in the 2016 Women's Olympic Marathon

David Cottrell and Michael C. Herron

*"I invested all I had and 300 meters before the finish line, I was next to Lisa. It was a magical moment that we could finish this marathon together. We did not think about what we were doing." – Anna Hahner*

At 9:30 a.m. on August 14, 2016, the women's Olympic marathon kicked off in Rio de Janeiro, Brazil, and 156 runners from 80 countries across the world left the starting line en route to their destination, 42.195 kilometers away. Two hours, 24 minutes, and 4 seconds later, Jemima Sumgong of Kenya would be the first to cross the finish line and take home gold; Sumgong was just 3 and a half minutes slower than her previous personal best time in the marathon.

Approximately 21 minutes later, twin marathoners from

Germany—Anna and Lisa Hahner—would cross the finish line together, holding hands and celebrating a personal victory. Although the Hahners would finish 81st and 82nd, respectively, well behind the winners of the marathon, Anna Hahner would describe their joint finish as a "magical moment."

The media quickly picked up the Hahner story as an image of the beaming twins, finishing hand-in-hand, captured a public audience. While many believed the twins' near-simultaneous finish was a reflection of Olympic spirit, not

everyone agreed with this rosy interpretation. The twins' happy facial expressions at the finish were portrayed as a bit contrived—smiling like Honigkuchenpferde (cookies in the shape of a horse) was the description offered by one editorialist—and the sports director of the German Athletics Federation, Thomas Kurschilgen, stirred up controversy when he suggested that the Hahners' photo-finish was no coincidence.

Kurschilgen averred that the twins slowed down deliberately to finish simultaneously and create a spectacle, and he justified

his charge by noting that the Hahners ran the Rio marathon more than 18 minutes slower than their personal best times before the Olympics. The Hahner twins denied Kurschilgen's accusations, perhaps not surprisingly.

What happened in the women's Olympic marathon in Rio, and how might we develop a statistical approach that assesses whether the Hahner twins' finish in the race was coincidental or intentional?

These two interpretations are clearly at odds. If the former, then the Hahners are to be celebrated and their finish treated as an expression of the spirit behind the Olympic games. If the latter, though, then the twins may have violated this spirit by not trying hard enough. It is perhaps too easy for us to write such a glib sentence—neither of us can fathom being able to complete a marathon anywhere in the vicinity of 2 and a half hours—but we nonetheless want to know what the data from the Olympic marathon tell us.

Among female Olympic marathoners, the Hahner twins were not alone in sharing familial ties. The Rio marathon also featured twins from North Korea, Kim Hye-song and Kim Hye-gyong, who posted identical times and finished 10th and 11th in the race, respectively. However, the Kim finish, unlike the Hahner finish, appears devoid of post-race controversy. Moreover, three triplets from Estonia competed in the Rio marathon, although only two, Lily Luik and Leila Luik, finished it, in 97th and 114th place, respectively. The third Estonia triplet, Liina Luik, recorded what is known as a DNF—"did not finish." Although our focus here is the Hahner twins, we also touch on the Kim twins and Luik triplets.

## Marathon Data and Our Research Design

For each participant who started the women's Olympic marathon, we know several things: personal best marathon time before the 2016 Olympic games; age; split times from the Rio marathon course at 5 kilometers, 10 kilometers, and so forth; and overall finishing time.

We cannot directly observe the effort that an individual put into the race, and we do not know why some runners have DNF results: Some runners may have injured themselves on the course and accordingly dropped out, and others may have stopped running, uninjured, in anticipation of an unsatisfactory result.

Of 156 marathon starters, 133 completed the race and 23 DNFed at various locations throughout the course. The overall DNF rate was thus  $23/156 \approx 0.15$ , and the relatively small sample size at our disposal yields a relatively wide 95% confidence interval for this rate, namely (0.098, 0.22).

Kurschilgen's accusation against the Hahner twins has two components: that these two women ran slowly *and* that they finished simultaneously. We suspect that Kurschilgen would not have expressed ire at the Hahners had they finished in first and second place in Rio, hand-in-hand and with wide grins. Thus, our investigation of the charges that Kurschilgen offered distinguishes between a slow finish and a simultaneous finish.

Our research design is twofold. First, we present visualizations that describe various features of the 2016 women's Olympic marathon; among other things, our visualizations feature differences between runners' Rio times and their prior personal best marathon times. The visualizations suggest that the

Hahner twins' pace in the marathon was slow, albeit not excessively so, but that their simultaneous finish was quite unusual given the twins' differences in abilities (and similarly for the Kim twins). We then turn to a regression-based simulation of the marathon, and our simulations reinforce what we observed in prior visualizations: that the Hahner twins did not run appreciably slowly, yet finished suspiciously close to each other. We return in our conclusion to Kurschilgen's claims about the Hahners and offer thoughts about their validity.

## Visualizing the Olympic Marathon

One way that we might assess whether a woman marathoner's Rio finish was unusual—or, say, whether two finishes were jointly unusual—is by comparing a runner's observed Olympic finishing time on August 14, 2016, with a measure of her underlying marathon talent. For the latter, we use prior personal best marathon times. We believe this is a natural measure of marathon talent, but it is not entirely free of complications. For example, personal best times are potentially confounded by the marathon courses on which they were set; some courses, like the Berlin marathon, are known for relatively fast times. In addition, personal best times may be confounded by conditions, such as weather, that vary across races.

Finally, personal best times may not capture Olympic race day idiosyncrasies that could affect individual runners. For example, a runner could have woken up with a minor cold on the morning of August 14, 2016. With these concerns in mind, we use a runner's half-marathon split time from the Olympic marathon as a

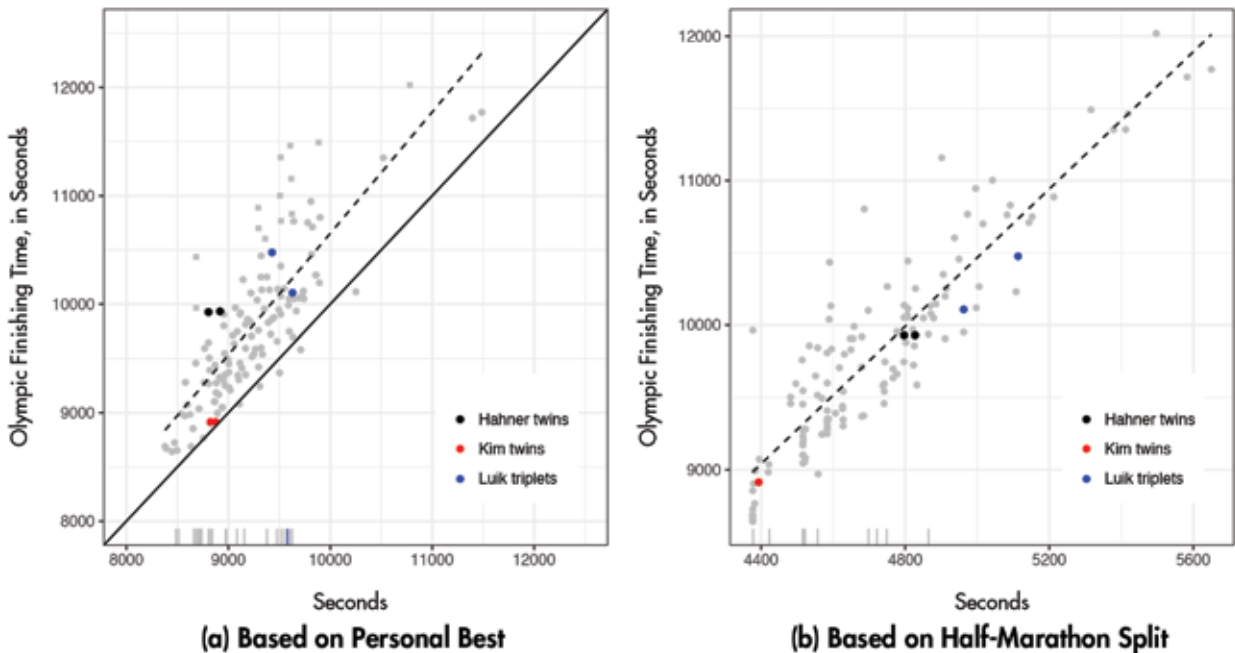


Figure 1. Olympic finishing times as a function of underlying marathon talent.

secondary measure of athletic skill in marathoning.

Figure 1 contains two plots describing how finishing times from the Rio women's marathon varied as a function of athletes' personal best times and half-marathon split times. The points in the plots are colored by twin/triplet status, and both plots contain dashed lines representing least squares regression fits. The marks along the horizontal axes in both plots indicate times, either personal bests or half-marathon splits, of runners who earned DNF results.

Considering first the relationship between Olympic finishing and personal best times, Figure 1a's solid 45-degree line is informative. Given the paucity of points (only five of them) below this line, it follows that the vast majority of Olympic marathoners ran slower in Rio than their prior personal bests.

The Hahner twins were definitely on the slow side, well above the 45-degree line, but a number of runners had even greater differences between their Olympic times and their personal bests. These runners are denoted with squares in Figure 1a, and there are 13 such symbols, highlighting approximately 10% of the racers who completed the Rio marathon. Moreover, the figure shows that two women had personal best times slightly faster than the Hahnners and yet finished after the two German women. Although Figure 1a suggests that the Hahner twins were slower than one would have expected given their previous best marathon times, it is not consistent with the accusation that they dramatically slowed down in the Rio marathon.

With respect to our second measure of marathon skill, Figure 1b shows that there was nothing abnormal about the Hahner twins'

overall finishing times, conditional on their half-marathon splits. As one might expect, a runner's time halfway through the course is a fairly strong predictor of her finishing time. Here we see that the relationship between the Hahner twins' half-marathon split times and their finishing times is similar to that of other Rio runners. The two points representing the Hahnners are in the middle of the distribution of points and therefore imply that the twins do not seem to have slowed down dramatically after they reached the halfway point of the Rio marathon. This is consistent with the aforementioned Figure 1a and inconsistent with accusations made against the Hahner twins; at least the part of the accusation that focused on their overall pace.

Another perspective on the extent to which the Hahnners' overall marathon finishing times were not unusual can be gleaned



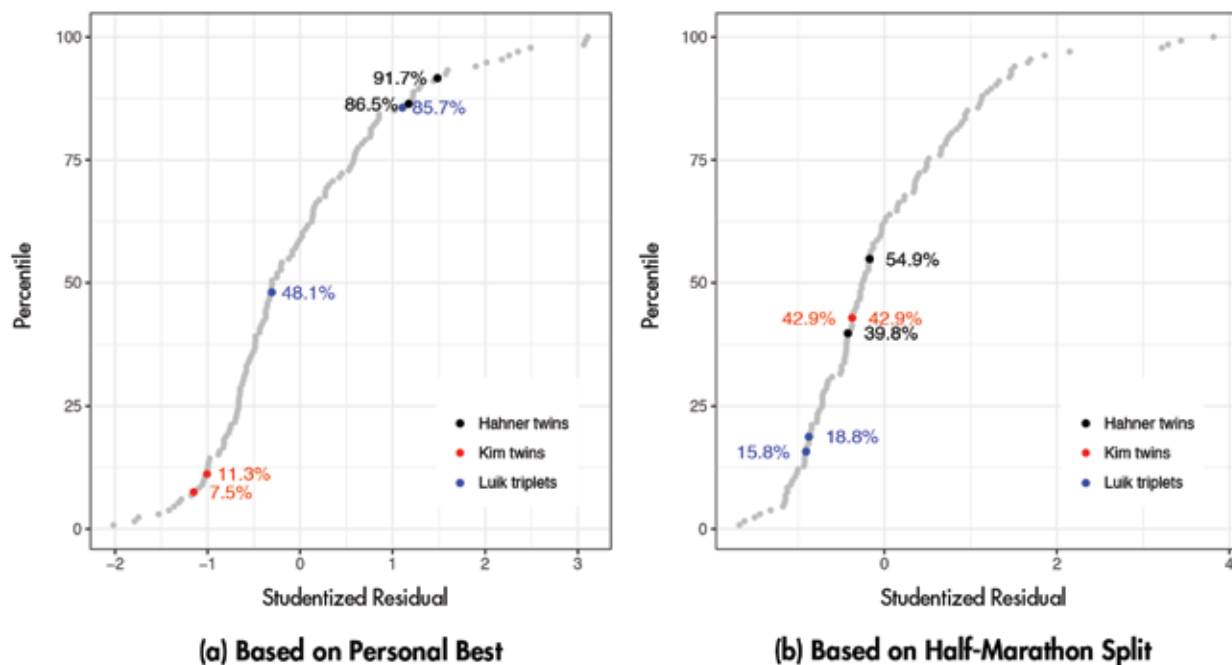


Figure 2. Studentized residuals from Olympic finishing time regressions.

from Figure 2. This figure presents cumulative distributions of the residuals from the regression models displayed in the two panels of Figure 1. In particular, we calculated the residuals from Figures 1a and 1b, Studentized them, and then arranged resulting Studentized residuals from least to most along the horizontal axes in Figures 2a and 2b.

The cumulative residuals depicted in the two panels of Figure 2 are consistent with our interpretation of the Hahner twins' Olympic finishing times as not particularly remarkable. With personal best time as a measure of marathon talent as in Figure 2a, the two Hahner residuals are in the right tail of the residual distribution; however, their locations are not extreme: One residual is located at the 87th percentile and the other at the 92nd.

While these two residuals are in the tail end of the residual

distribution, they are not major outliers that would lead us to think that the Hahner twins' marathon finishing times were extremely slow. Moreover, the residuals for the Kim twins are similarly unremarkable; these two residuals are in the left tail of the residual distribution, indicating that the Kims ran faster than one would have expected.

Finally, if one relies on half-marathon split times as measures of marathon talent, as in Figure 2b, similar conclusions follow. Neither Hahner twin had a finishing time that was particularly unusual given her half-marathon split, and this applies to the Kim twins as well.

If the Hahner twins did not slow down excessively, might they have run somewhat strategically at the end of the Olympic marathon to generate a simultaneous finish? This visualization speaks to this question.

The personal best times of the Hahner twins were 115 seconds apart and their official finishing times were separated by 1 second. Is such a 115 to 1 compression typical among pairs of runners? Have other pairs of Olympic marathoners had a difference between personal best times of 115 seconds apart and, if so, how close were their finishing times?

Of the 133 marathon finishers, there are  $\binom{133}{2} = 8,778$  pairs of runners. Of these and ignoring the Hahner twins, 10 had exactly a 115-second gap in personal best times. Differences in finishing times of these 10 pairs, in seconds, are: 36, 93, 172, 319, 379, 459, 552, 671, 675, and 739. In other words, of all pairs of runners in the Rio marathon who had a personal best difference that was equivalent to the Hahner twins' difference, the twins had the greatest compression based on finishing time.

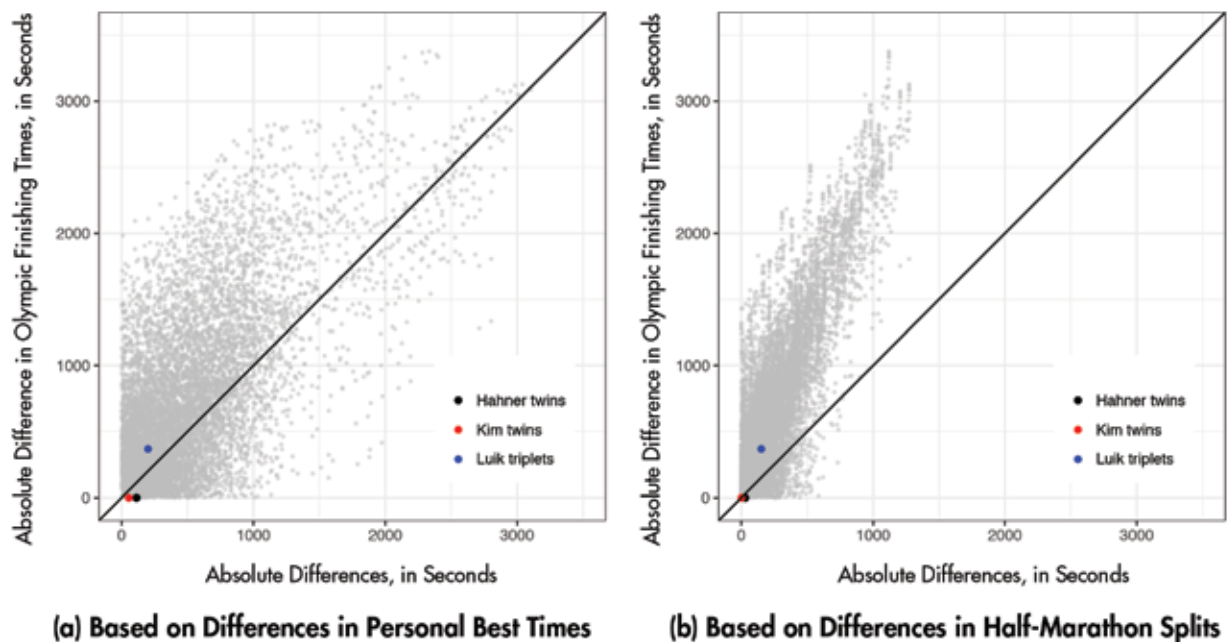


Figure 3. Pair-wise differences in Olympic finishing times and differences in marathon talent.

We can generalize this result by looking at all pairs of runners in the marathon. For all 8,778 pairs of 133 finishers, Figure 3a plots differences in finishing times against differences in personal best times, and pairs of twins/triplets are identified by the same color scheme used earlier (recall that only two of the Luik triplets finished the Rio marathon).

Consider first the Hahner twins. The two German women are effectively located on the horizontal axis because their difference in finishing times is 1 second. However, there are many points above the Hahnners' black dot, and this shows that, conditional on an approximate 115-second difference in personal best times, most marathoner pairs did not have close finishing times like the Hahnners. Some pairs of runners with around 115-second personal best differences had finishing time differences of 1,000 seconds, i.e., more than 15 minutes.

The points in Figure 3a are not independent, but they provide a rough sense of the dispersion in finishing time differences between runners that one might expect, conditional on differences in personal best times.

Thinking about the accusations leveled against the Hahner twins, Figure 3a suggests that Anna and Lisa Hahner did indeed run with an eye on each other. In fact, the same can be said of the Kim twins, who ran seemingly in lockstep throughout the entire Rio marathon. The North Korean twins had a personal best difference of 53 seconds and a finishing time difference of literally zero seconds. Beyond these twins, eight pairs of Rio runners had a 53-second personal best difference and resulting finishing time differences of 9, 51, 228, 340, 352, 571, 662, and 751. As in the Hahner case, the Kim twins compressed their finishing times—meaning that they finished

with less time between them than the difference in their personal bests—more than any other pair of runners with similar personal best differences.

Similar conclusions follow from Figure 3b, which plots pair-wise differences between Olympic finishing times and half-marathon split times. Namely, many pairs of runners had similar differences in half-marathon times as the Hahner and Kim twins, but the vast majority of these pairs did not have close finishing times.

Figure 4 describes each Olympic runner's status at various split times on the marathon course. Each dot in the figure—colored as before—depicts a recorded split and the number of seconds each runner was behind the race leader at the time. There are more dots at earlier splits due to subsequent DNFs.

The Estonian triplet DNF before the half-marathon split is

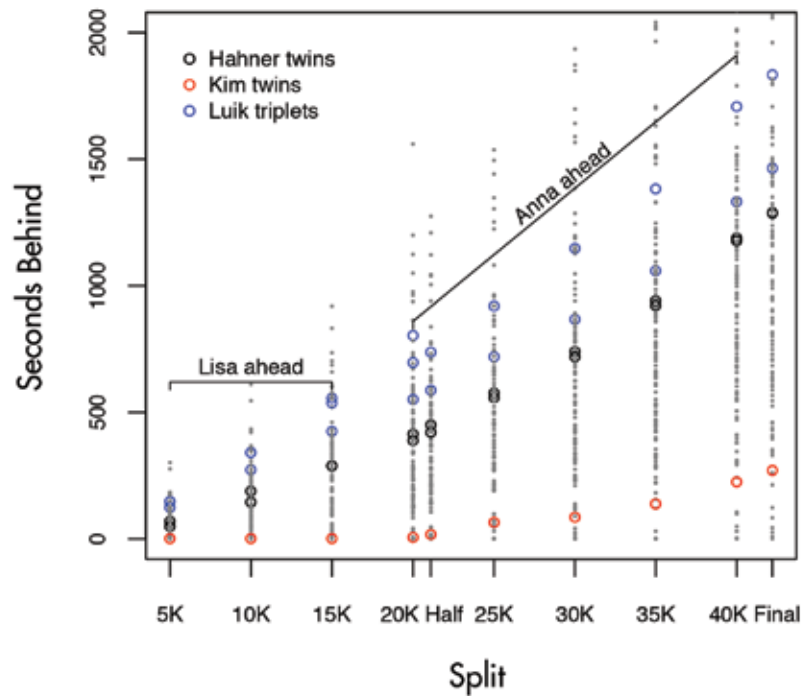


Figure 4. Runner status by split.

evident in Figure 4, which also shows that Lisa Hahner was ahead of her sister through 15 kilometers. The figure contains two red dots representing the North Korean Kim twins, but this is not visually apparent because the Kim twins had identical split times during the entire marathon. This explains using the term “lockstep” to refer to the Kim twins’ marathon pace.

## Simulating the Marathon

Our visualizations shed a fair bit of light on the Hahner twins’ performance in the 2016 women’s Olympic marathon. In the interest of increasing precision, this question arises: If we take into consideration the twins’ similarities in marathon talent and natural variation in marathon finishing times, what is the probability that they would finish the Rio race at roughly the same time and/or sequentially?

To answer this question requires knowing the counterfactual distribution of potential marathon finishing times that would have occurred had the Hahner twins independently (in particular, of each other) and repeatedly run the Rio marathon, holding constant marathon conditions, the abilities of other runners, and so forth. Access to such a distribution would establish the set of potential outcomes that could have occurred on August 14, 2016, and we could in principle use this distribution to determine the likelihood that a simultaneous finish by the Hahnners, or at least a near-simultaneous finish, occurred by chance alone.

If these twins rarely finish the marathon together in such a counterfactual world, then one might be skeptical that their observed finish occurred without some degree of coordination.

Unfortunately for us, but fortunately for the race participants, it is

not possible to rerun the women’s Olympic marathon to establish a distribution of potential race outcomes for the Hahner twins. However, we can attempt to simulate this distribution by estimating the distribution of every other runner’s finishing time, conditional on marathon talent, and then drawing from this distribution to calculate the likelihood that, for example, Lisa and Anna Hahner finished the Rio race simultaneously.

To estimate the conditional distribution of each Rio final result, we assume that each runner’s marathon time  $Y_i$  is distributed normally with a mean that is a function of the runner’s marathon skill, which we assume is a linear combination of her personal best marathon time  $X_i$  and her age  $Z_i$ :

$$Y_i | X_i, Z_i \sim N(\beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i, \sigma)$$

where  $N(\cdot, \cdot)$  denotes a normal distribution. We estimate  $\beta_0, \beta_1, \beta_2$ ,

and  $\sigma$  using ordinary least squares on finishing Rio marathoners. We exclude the Hahner/Kim twins and Luik triplets from the sample so our estimates are not affected by the twin/triplet finishes which, theoretically, could reflect runner coordination.

For a simulated marathon, we draw a runner's time from an estimated distribution and condition on the runner's personal best marathon time and age. Once a race is simulated for all runners, twins and triplets included, we record both the time between Anna and Lisa Hahner's simulated finishes and the difference in their simulated ranks. We then simulate a new race—drawing a new set of finishing times—and record the same quantities. These are the steps in the simulation.

1. Ignoring twins and triplets, estimate a linear model with least squares that predicts a runner's finishing time  $Y_i$  based on her pre-Olympic personal best time  $X_i$  and her age  $Z_i$ .
2. Extract the resulting coefficient vector, estimated covariance matrix, and estimated regression variance from this model.
3. For each simulated race, draw intercept and slope estimates  $\tilde{\beta}_0$ ,  $\tilde{\beta}_1$ , and  $\tilde{\beta}_2$ , respectively, from a multivariate normal distribution with mean equal to the previously estimated coefficient vector and covariance equal to the previously estimated covariance matrix.
4. For each runner, draw an error  $\tilde{\epsilon}$  from a normal distribution with mean zero and a standard deviation equal to the standard

deviation of the original regression model's residuals. This step requires 156 draws from a normal distribution; one draw per marathoner.

5. Predict each runner's final result by combining the randomly generated beta coefficients and individual error terms,  $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 X_i + \tilde{\beta}_2 Z_i + \tilde{\epsilon}$ .
6. Eliminate each runner from the simulated race with a probability equal to the observed fraction of marathoners who did not finish the marathon.
7. Repeat above steps 10,000 times.

As these steps illustrate, our simulation repeatedly draws random coefficient vectors, and this captures uncertainty in what we know about the relationship between runner talent and age and runner finish times. In addition, for each simulated race, our simulation draws random disturbances for each runner, conditional on the original estimate of regression variance; these disturbances capture variability in runner finishing times, notwithstanding age and underlying marathon talent. Importantly, the disturbances that we draw are independent across runners. Consequently, for each simulated race, the finishing order among runners will vary.

From our simulations, we generate intervals that describe the extent of the variability in marathon finishing times. For example, in 95% of the simulations in which Anna Hahner completed the marathon, her finishing time was between 8,369 seconds and 9,999 seconds. This is consistent with Anna's observed finishing time in Rio, which was 9,932

seconds. Nonetheless, Anna's finishing time was on the slower end of this interval. In fact, our simulations estimate that, given Anna's personal best marathon time and her age, she would be expected to finish the Rio marathon in 9,181 seconds, which is slightly more than 12 minutes faster than her actual time.

This is similar to her sister's result. Lisa Hahner's corresponding 95% interval is 8,507 to 10,143 with a mean of 9,319. Her actual finishing time was 9,933 seconds.

The bottom line here is that our simulated marathon finishes are consistent with the Hahner twins' finishing times in that their finishes were inside the bounds of traditional 95% intervals. And note that the regression model underlying the simulations was estimated without the Hahner twins (and same for the Kim twins and Luik triplets). According to our simulation, then, both Hahner twins did not run appreciably slowly, conditional on personal best times before the Rio Olympics and age.

With an eye on the matter of simultaneous finishing, Figure 5 contains two histograms based on simulated race results. Figure 5a is a histogram that shows the distribution of absolute differences in Hahner twin finishing time where differences are grouped in 30 second bins; counts for the various bins are denoted by the vertical lengths of the bars. Figure 5b is similar, but depicts the distribution of the absolute differences in Hahner twin rankings. Differences are grouped as single units, ranging from no runner between the twins to nearly 120 runners between them.

The histograms in Figure 5 raise questions about the credibility of Anna and Lisa Hahner's story and, in particular, suggest that a simultaneous finish in the Rio marathon

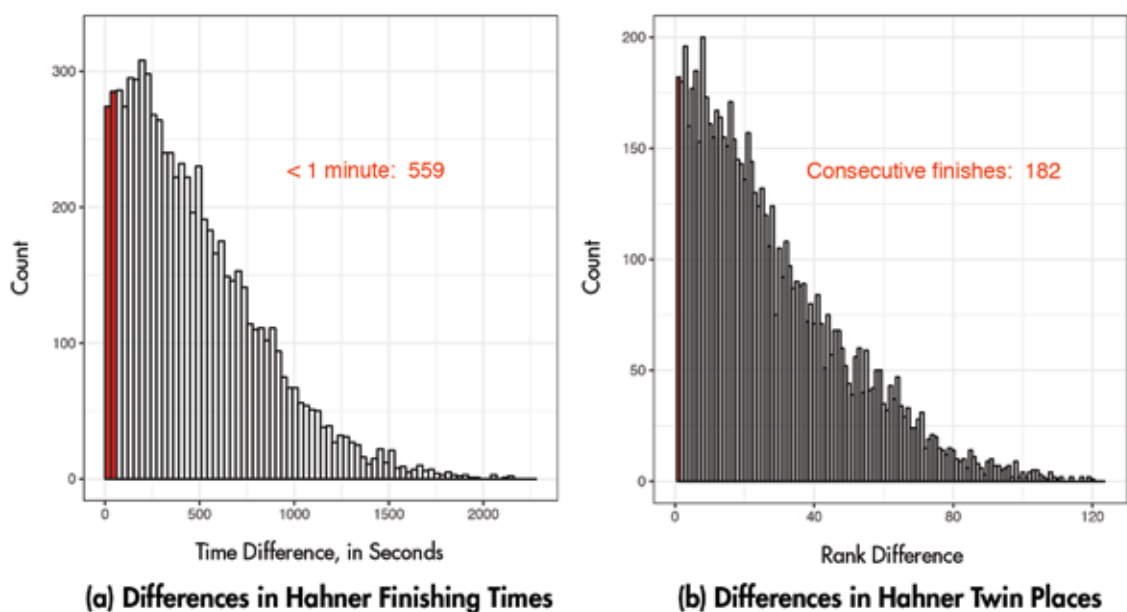


Figure 5. Distribution of Hahner twin results in 10,000 simulated marathons, based on personal best times.

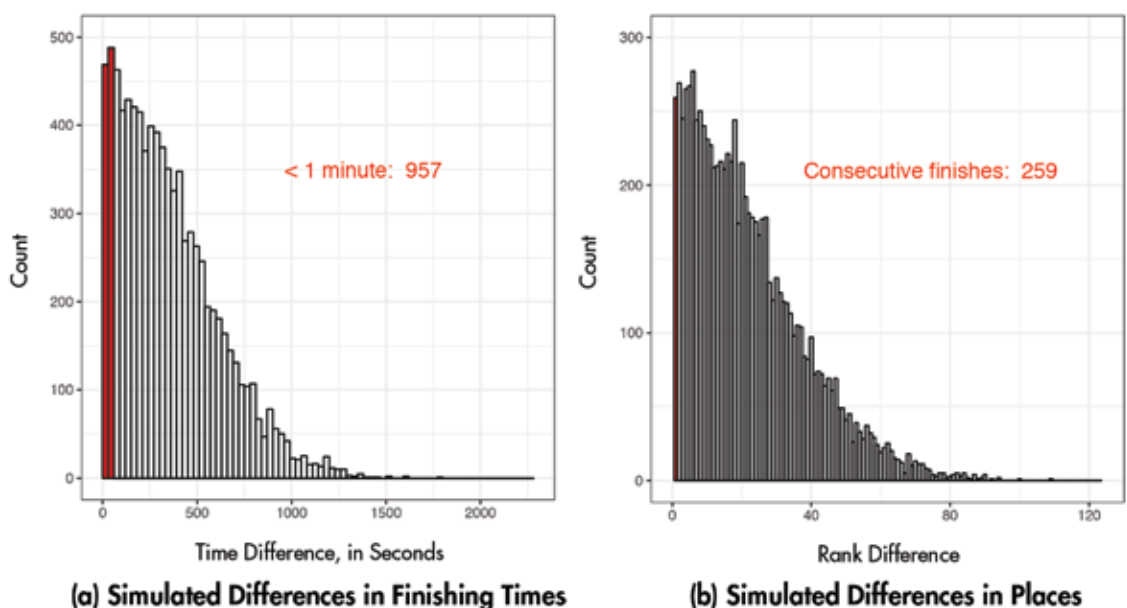


Figure 6. Distribution of Hahner twin results in 10,000 simulated marathons, based on half-marathon splits.

would be very rare if Anna and Lisa had run independently. For example, in fewer than 300 of 10,000 simulated races did Anna and Lisa Hahner finish within 30 seconds of one another, and in fewer than 600 did the Hahner twins finish within a minute of each other. The histogram area

associated with this latter result is depicted in red in Figure 5a.

Moreover, the Hahner twins finished in consecutive rank in fewer than 200 of 10,000 simulated races; the red zone in Figure 5b presents this visually. The close finish that was observed in Rio, where Anna and Lisa crossed

the finish line one after the other, would have been highly unlikely if the two German twins had raced independently of each other.

Parallel to our prior analyses, we repeated our simulations using half-marathon splits as the predicting variable in our simulations; results are in Figure 6. While this



use of marathon splits reduces the variation in the predicted outcome of each runner and therefore reduces the expected distance between Anna and Lisa Hahner, even with half-marathon split as a measure of ability, it is still quite rare for the twins to finish simultaneously or consecutively.

The way that we handled DNFs in our simulations is notable. As our description indicates, we assumed that DNF probabilities are the same for all runners and that the likelihood of a DNF is not a function of a runner's anticipated marathon finishing time. The rug marks in Figure 1a suggest that runners with better personal best times may be more likely to DNF than other runners, all things equal. We suspect that this occurs because some better runners may expend excessive energy trying to achieve a good result in the marathon and in so doing, injure

or exhaust themselves; lesser runners, in contrast, may be content to finish respectably.

Regardless of the validity of this conjecture, Figure 1a shows that the Hahner twins are representative of the sort of runners who DNFed in Rio. Since our simulated Hahner statistics are conditioned on both Hahner twins finishing, it follows that they are conservative.

The fact that both women finished the marathon was notable in and of itself and, by discounting the possibility of a Hahner DNF, we are giving the benefit of the doubt to the Hahner twins.

## Conclusion

Anna and Lisa Hahner's near-simultaneous finish in the 2016 women's Olympic marathon in Rio elicited a controversy in that the German twins were accused of deliberately slowing down and finishing next to each other to generate media attention. They denied this, not entirely surprisingly.

Of one perspective on the marathon that is based on visualizations and simple calculations, and a second that draws on a simulation, both have the same implications. In a global sense, the Hahner twins did not slow down appreciably during the Rio marathon. Their times were not fast, but they were within reason for runners of the Hahners' abilities and age. Locally, though, we find that the Hahners' finish probably was, in fact, contrived. Their finish—consecutive, with 1 second between the two women—was a rather low-probability event. Compared to their differences in talent, that is, the Hahner twins

difference in finishing times was unusually compressed. This is evidence that their finishing had elements of intentionality—perhaps at the last minute, but intentionality nonetheless.

Our goal is not to speak to whether the Hahner twins should or should not have enjoyed what some might call an artificial moment of Olympic glory. Neither was in contention for a podium finish in Rio, and compared to the doping allegations that presently surround endurance sports in general, what the Hahners appear to have done seems relatively tame. Still, it might have behooved them to have been a bit more open about their end-of-race tactics. To this end, it is clear how a simple data analysis can shed light on claims about racing results.

Should the 2020 Summer Olympics again feature marathon-ing twins and triplets, we look forward to a comparative analysis using the techniques illustrated here. ■

## Further Reading

<http://www.telegraph.co.uk/olympics/2016/08/17/german-twins-criticised-for-finishing-olympic-marathon-fun-run-h>.

<https://www.nytimes.com/2016/08/17/sports/olympics/twins-finish-marathon-hand-in-hand-but-their-country-says-they-crossed-a-line.html>.

<https://www.welt.de/sport/olympia/article157669264/Das-falsche-Laecheln-der-deutschen-Lauf-Zwillinge.html>.

## About the Authors

**David Cottrell** is a lecturer in the Department of Government at Dartmouth College and was a postdoctoral research fellow in Dartmouth's Program in Quantitative Social Science when he completed work on this article. He teaches courses on the application of data analysis and statistics in the social sciences and is currently involved in research that leverages computational approaches to study the effect of gerrymandering.

**Michael C. Herron** is William Clinton Story Remsen 1943 Professor of Government and chair, Program in Quantitative Social Science, at Dartmouth College and was a visiting scholar at the Hertie School of Governance, Berlin, Germany, when he completed work on this article. He has taught courses in applied statistics at Dartmouth since 2004. At present, he is researching relative age effects in professional football and the extent to which Americans vote infrequently, thus exposing themselves to the risk of being removed from registered voter pools.

# The Point(s)-After-Touchdown Decision Revisited

Harold Sackrowitz

In American football, the game and the clock are temporarily stopped after one team scores a touchdown. Then that team has the opportunity to score either one or two extra points. The team can get one additional point by kicking the equivalent of a short field goal. In a more-difficult option, the team can get two additional points by scoring the equivalent of a short touchdown. After this attempt, the game resumes and the clock restarts.

Regardless of what occurred during this extra point phase, the team that scored the touchdown kicks off to the other team to resume the game. Of course, the number of points scored during these bonus situations can have a big impact on winning or losing.

In 2000, I wrote an article for *CHANCE* magazine about optimal strategies for making these Point(s)-After-Touchdown (PAT) decisions. In that article, I described how dynamic programming methods could be used to generate an extra-point strategy table. Before the start of the 2015–16 season, the NFL implemented a rule change to make the one-point conversion (kick) attempt more challenging. (The two-point conversion rule remained the same.) This could change, to an unknown degree, the balance between the two options. Now, after three years under the new rule, it appears be a good time to review what has happened since my last article and



the substantive implications of the new rule.

The two-point conversion option has been used in college football since 1958 and in professional football since 1994. In 1998, I became aware of something called “the chart”—a chart constructed to indicate the proper strategy after a touchdown. It listed all possible point differentials at the moment after a touchdown and indicated whether a team should try for one or two

extra points. Its development is credited to Dick Vermeil when he was on the UCLA coaching staff, and every college and professional coach seemed to have (and still have) a copy.

What I found most curious was that there was no mention of how much time was left in the game when a decision was to be made, yet coaches did seem to recognize that a decision that was correct late in a game might not be correct early in a game. In fact, most coaches did

not like to try for two before the fourth quarter.

I was teaching a probability class at about that time and thought that considering some extra point situations might make for interesting class examples. One such example is the case of a team being behind by 14 points when they get the ball late in the game. The coach also believes that they will have, at most, two more possessions in the game. Thus, they must score two touchdowns while their opponent does not score at all. With that mindset, how should they plan their extra point strategy?

For this case, Porter did some calculations published in a 1967 *American Statistician* article. He pointed out that, if the probability of a successful two-point conversion was at least .382, that team should try for two points if they score a touchdown on their first possession. If successful, they would potentially be in position to win outright with another touchdown. If not, another two-point try could create a tie.

This type of question ultimately led to my article in *CHANCE* in 2000. My tables were based on the score differential, number of possessions remaining in the game, success probabilities of one- and two-point conversion attempts, and scoring rates (for field goals and touchdowns).

To my surprise, that investigation was mentioned in articles in both *Sports Illustrated* and the *New York Times*. It even led to a couple of sports radio talk show interviews. Shortly after the *Times* article appeared, I was contacted by Ernie Adams of the New England Patriots. At that time, they were using a version of “the chart” that they had modified based on intuition. We had several conversations to discuss the elements of

what went into my calculations. He asked relevant questions—he wanted to know which factors had the greatest impact. In particular, he asked in which situations did I feel strategies were clear-cut and which were gray areas. I was both surprised by his phone call and impressed by his understanding.

Based on these early experiences, I began to think that my results would generate a great deal of interest. This was not the case. On most Mondays during each NFL season, I review the box scores of the weekend games. If I see a game that was lost when following my table might have saved it, I would sometimes write to the coach with the following offer: Have someone on your staff speak to me about the basis and construction of my tables to see if you think it is sensible and perhaps worthwhile to use them. With one special-case exception, that offer has either been graciously turned down or ignored for the last 17 years.

The one exception occurred in 2009, when Greg Schiano was coach of the Rutgers University football team. A fellow faculty member who was friendly with both of us mentioned my work to Schiano. He had enough quantitative curiosity to invite me to give a presentation to him and some of his staff. I explained that the method essentially examined every possible way a game could play out and gave weights to the different scenarios. I showed them many possible strategy tables with differing success rates. I gave them examples of past Rutgers games where going for two would have been a better choice than what they had done. In one example, I showed them a printout of about 900 ways a particular game could have played out so one could see

that going for two was preferred in most scenarios.

After 90 minutes of discussion with questions and answers, they were convinced that my tables could supply some valuable information. Even after Schiano became coach of the Tampa Bay Buccaneers in the NFL, he considered my tables in his planning.

Despite all the current publicity about how the world of sports will be turning to statisticians, that has not been my experience. The explosion has really been in the ability to collect data—teams can now keep a record of what every player was doing on every play. It is true that, in recent years, some teams have added people with very strong statistical credentials. However, the types of strategy analyses of current interest to coaches are well within the grasp of any bright person who has taken an introductory statistics course and has some computer skills. Teams are particularly interested in tendencies appearing in these large data sets.

## Some Specifics

When I began my study in 1998, the ball was placed at the two-yard line by PAT rule and the team could try for either one or two points. At that time the success rate for a one-point kick was 0.987 and the two-point conversion success rate was 0.39. Since that time, both yearly success rates have followed an upward trend. The two-point conversion success rates have more yearly variation since there are typically fewer than 100 attempts in any one year.

On the other hand, there would be more than 1,100 kick attempts per year. In the three years (2012–2014) before the change, the two-point success rates wavered between 0.47 and 0.50, but the successful kick rate for the 2014

season was up to 0.993. The NFL recognized that the kick was viewed as “automatic” by fans and possibly had become less interesting than the pregame coin toss.

This precipitated a rule change by the NFL. Beginning with the 2015–16 season, the ball was to begin at the 15-yard line for an attempted kick instead of the two-yard line. This meant that the actual kick would be about 33 yards. As the 2015–16 season approached, the rule change received considerable media attention. The consensus opinion was that field goal kickers were so good from that portion of the field that the extra 13 yards would not matter. Coaches did not foresee it affecting their decisions and were not planning to rethink their PAT strategies.

Although it did make the extra point kick more interesting to watch, this rule change was deemed minor. However, from a mathematical point of view, the NFL could not have done much better even had they known what they were doing.

## Why a Model is Needed

It is easy to find dozens of articles on the Internet that comment on strategy under the new rule. Commenters know the PAT success rates under the new rule and try to make intuitive arguments based on them. That is not good enough, though, especially before the fourth quarter.

Coaches decide to try for two points, almost exclusively, in situations when their team is either ahead by one point, ahead by five points, behind by five points, or behind by two points. This is not likely to ever change (although it probably should). The vast majority of two-point attempts occur in the fourth quarter when fewer

opportunities remain for teams to score points. These decisions are typically based on what the coach envisions might happen in the short time remaining. People who follow football are used to points being scored in groups of three and seven. That is what they are thinking when a PAT decision must be made. In most cases, neither experience nor probability is used (or required) toward the end of the game—just simple logic.

The 2016–17 season’s playoff games provided good examples of the “behind by two points” case. In the second round of the playoffs, Kansas City scored a touchdown against Pittsburgh and was behind by two points, with 2:43 remaining in the game. In another playoff game, Miami scored a touchdown against Green Bay and was behind by two points, with 4:08 remaining in the game. Both attempted a two-point conversion. Every other coach (and probabilist) would have done the same thing.

The reasoning is that being behind by two is no worse than being behind by one at that point in time. Equivalently, a failed two-point attempt is no worse than a successful one-point attempt. On the other hand, a successful two-point try would tie the game.

This argument is so compelling that although Kansas City committed a foul during their attempt, they still tried for two despite being penalized 10 yards. They failed and lost the game. Miami’s two-point attempt was successful, but they lost anyway. These negative results would not (and should not) change the coaches’ future decisions in similar settings.

In contrast is the Houston versus New England playoff game. There, the Houston team scored a touchdown and also found themselves behind by two points, but this time, there were more than 40

minutes remaining in the game. My tables suggest that trying for two would have been optimal. Instead, Houston successfully kicked the extra point. Ultimately, though, Houston lost 34–16, indicating that neither their choice nor my choice would have made any difference to the outcome. In effect, nothing about PAT strategy was to be learned from that game.

The problem is that, for any fixed point differential, it becomes more and more difficult to use the approach of trying to visualize what might happen as time remaining increases. These situations cannot be resolved satisfactorily (certainly not before early in the fourth quarter) in an empirical way. An empirical approach would require each possible PAT situation (score differential and time remaining) to occur many times. Furthermore, both actions (try for one or try for two) would have to be taken many times in each setting.

Finally, it would have to be clear by the end of the game which decision would have been correct. One cannot simply record whether the game was won or lost. There are actually very few games, in an entire NFL season, in which the choice of a PAT action taken before the fourth quarter can actually affect the game result. The Houston–New England game above is typical of what happens by game’s end. Think about how many games would have to be played to resolve this situation empirically.

Another roadblock is that coaches are not going to vary their decisions just to make their games part of a controlled experiment.

In addition to all that, the rules changed only three years ago. Even the most experienced of football people could not have amassed enough relevant information to develop reliable intuition.



I do often wonder about what coaches base their decisions on in non-obvious cases. They seem to go through phases in which aggressiveness goes in and out of style. In each of the last three years, there were between 88 and 112 two-point attempts. This followed a number of years when the number of attempts were in the 50–60 range. In 1999 and the early 2000s, when I first began looking at the issue, the number of attempts were in the 80–100 range. Back then, teams were not nearly as proficient at converting two-point tries as they are now.

It is true that, in some years, there are just more games that are close near the end than in other years. In some years, score differentials also arise in more games that are conducive to two-point attempts. I do not think that is enough to explain it.

On occasion, there will be a game with a PAT decision that actually seemed debatable to the media. If I see such a game when checking box scores, I go to the website of the team's local newspaper to read accounts of the game. I do this because questions about the decision can come up in the coach's post-game interview. Coaches' responses are in the spirit of "it was early in the game and a lot could still happen"; "if I had it to do over again, I would do the same thing"; "I didn't want to get beat by two field goals"; "that is what the chart says"; "I wanted to be more aggressive"; and "we didn't come here to tie, we came to win." I have never heard the words "in my experience..." or "it was the percentage play." Generating tables like mine would let a coach make a much-more informed decision.

## How It Works

Since my table appeared, others (such as Krasker on *footballcommentary.com*) have done similar, more-elaborate and ambitious things. Such tables (as well as the original "chart") can be found by surfing the web. A data-driven approach is also gaining some interest based on so-called win probabilities. The current version is due, mainly, to Burke on *advancedfootballanalytics.com*.

First, you input all sorts of variables such as current score, time remaining, time outs remaining, believed team strength, etc. Then you are told the proportion of times that teams in that situation have gone on to win the game. For use in PAT strategy, it has a dynamic programming flavor.

After scoring a TD but before the PAT attempt, we know that team will add either zero, one, or two points. Suppose we know the "true" win probabilities associated with each of these possible outcomes. Then, if we also knew the "true" success rates for one- and two-point conversion attempts, we could come up with a strategy.

Of course, these win probabilities are based on results of teams that were not following optimal strategies. The fact is that no table can give a definitive strategy answer in all situations. There is no way that a coach's knowledge can be totally removed from the equation, and there are too many unknowns without good ways to estimate them.

The dynamic programming model presented in *CHANCE* previously (2000) requires the input of the success rates of both one- and two-point conversion attempts. It also requires the input of field goal and touchdown scoring rates of the two teams. All these can be thought of as initial conditions. It then yields a table that, based on

the score differential and number of possessions remaining, indicates the action that will maximize the probability of winning. None of the initial conditions or even the number of possessions remaining can be known exactly. It is not even clear that they can be estimated reliably.

It would not be correct to pool data for all teams from previous seasons—different teams have different scoring rates. In fact, personnel changes often make individual teams perform quite differently in successive seasons. Perhaps some of these quantities can be estimated by the second half of the season. Certainly the true individual two-point success rates cannot be estimated reliably. It is unusual for any one team to attempt more than six in an entire season.

How can this method help? Fortunately, the initial conditions do not always have to be known with much precision. This is crucial to making the tables into a valuable guide. It turns out that large portions of the optimal strategy tables are not particularly sensitive to changes in the initial conditions. If this were not the case, the practical value of this approach would greatly diminish.

The exception to this phenomenon is when the order relationship between the one-point success rate and two times the two-point success rate is reversed. This situation is addressed later. It is always essential to review a variety of tables based on different initial conditions and recognize their common regions. As an example, we compare the following two sets of initial conditions. For convenience in presentation, we simply assume that all teams have the same scoring rates.

**Initial Condition A:** 99% and 42% success rates for one-point and two-point conversion attempts



**Table 1—Optimal Strategy Chart for Initial Conditions A**

Behind by	Total Possessions Remaining in the Game																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
10			2	2	2	2	1*	2	1	1*	1	1*	1	1	1	1	1	1	1	1	1
9					2	2	2	2	2	2	2	2	2	2	2	2	1*	1*	1*	1*	1
8			2	2	2	2	2	2	2	2	2	1*	1*	1	1	1	1	1	1	1	1
7			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6			1	1	1*	1*	1*	1*	1*	1*	1*	1*	1*	1*	1	1*	1	1	1	1	1
5			2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1*	1	1	1
4			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3			1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ahead by																					
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1		2	2	2	2	2	2*	2	1	2	1	1	1	1	1	1	1	1	1	1	1
2		1	2*	2*	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1
3		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4		2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5		2	2	2	2	2	2	2	2	2	2	2	2	2	1*	1*	1*	1*	1	1	1
6		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9				1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10				1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Taking the opposite actions for all starred entries results in the optimal strategy chart for Initial Conditions B. 1 means try for one point; 2 means try for two points; blank means it does not matter. Note that a typical NFL game has 4-6 possessions per quarter.

respectively, with teams scoring field goals and touchdowns on 13% and 21% of possessions respectively.

**Initial Condition B:** 99% and 47% success rates for one- and two-point conversion attempts respectively, with teams scoring field goals and touchdowns 19% and 30% of possessions respectively.

A sense of this phenomenon can be seen in Table 1 (it is abridged to make it easier to focus on specific

point differentials). Table 1 shows the optimal strategy for Initial Conditions A. In addition, we have indicated with stars those entries where the optimal strategy under Initial Conditions B would require the opposite action. The two sets of strategies are very similar despite the difference in initial conditions. The only dramatic difference is in the behind-by-six row. Of particular interest are the

rows for behind by five and behind by two. Even here, there is very little difference between Initial Conditions A and B.

Thus, there are many situations where a coach could confidently make a decision, and there are others in which he would have to invoke his sense of how the game is progressing. For example, when behind by five, a team could comfortably, by my tables, go for two

any time in the second half. On the other hand, when ahead by five, the coach might want to wait until the middle of the third quarter. Here a coach's input is needed for a final decision. Even the number of possessions would typically only be known to be in some range. That is often enough to identify the optimal strategy. Otherwise, the coach would have issues to resolve: Will he be using a hurry-up offense, will the other team be using a time-consuming offense, etc.? In a typical game, there are approximately four to six possessions per quarter.

## Ramifications of the New Rule

Consider any new rule that makes the one-point conversion more difficult but leaves the two-point conversion alone. Next, think about any game scenario in which going for two had been the preferred choice under the old rule. That would create the expectation to be even more inclined to go for two under the new rule in that same scenario.

Using the data from these past three years to compute optimal strategies bears this out to a surprisingly large degree. Furthermore, the data lead one to an important end-of-game issue, not present under the old rule and in need of attention.

Let  $p_1$  denote the probability of a successful one-point conversion (kick) attempt and let  $p_2$  denote the probability of a successful two-point conversion attempt. Historically, when people have used intuition to venture quick PAT opinions, it has been implicit that  $p_1$  is approximately 1 and  $p_2 < 0.5$ ; in effect, that  $p_1 > 2p_2$ . Trying to kick for the one extra point has always been the default action in the NFL. The average value of a 1-point attempt is  $p_1$  points while the average value of a two-point attempt is  $2p_2$  points. If

$p_1 > 2p_2$  then, in a sense, a one-point attempt is worth more than a two point attempt. In that case, if a team always attempted a one-point conversion, it would maximize the total number of points it expected to score.

On the other hand, if  $p_1 < 2p_2$  were the case, then the opposite would be true.

Determining an optimal strategy is challenging because, unfortunately, maximizing the average number of points scored is not the same as maximizing the probability of winning. However, the math does say that, for a while at the beginning of a game, they can be the same. In fact, if  $p_1 < 2p_2$ , then trying for two in the very early part of a game might be optimal.

The combined success rates for the last three years—the first under the new rule—was a pleasant surprise. Based on 3,677 one-point attempts and 300 two-point attempts, the success rates were  $p_1 = 0.939$  and  $p_2 = 0.47$ . This is, almost exactly,  $p_1 = 2p_2$ . A closer look is informative.

During the 2015–16 season, the first under the new rule, success rates were  $p_1 = 0.942$  and  $p_2 = 0.5$ . The 2016–17 season success rates were  $p_1 = 0.934$  and  $p_2 = 0.491$ . However, the 2017–18 season success rates were  $p_1 = 0.942$  and  $p_2 = 0.409$ . This is an extremely low two-point success rate. An even closer look reveals that in the first half of the season, the success rate was  $14/45 = 0.31$  and  $22/43 = 0.51$  in the second half.

I cannot explain what happened in the first half of the season. It does show how hard it is to get a reliable point estimate of the true  $p_2$  that a coach could be confident in.

For the first two years, there was a combined  $p_1 = 0.938$  and  $p_2 = 0.495$ . If these numbers were true indicators of what success rates had become, then dramatic

changes in PAT thinking should ensue. Table 1 is representative of optimal strategies for initial conditions when  $p_1 > 2p_2$ . Now compare that to Table 2, which uses the first two years' rates of  $p_1 = 0.938$  and  $p_2 = 0.495$ .

This table is overwhelmed by twos. Of special interest is the "ahead by 6" line of Table 2. We see that with 20 possessions remaining, the optimal choice is to try for two. This would be the situation when the first touchdown of the game is scored very early.

An intuitive way to think about the relationship between  $p_1$  and  $p_2$  is as follows. When  $p_1 > 2p_2$ , the optimal strategy attempts to give a team an additional opportunity to avoid a loss in regulation time. For example, a team coming down to their last possession could be behind by four points, needing a touchdown to avoid a loss. Had they made a two-point conversion at some earlier time when it seemed appropriate to try for it, they would be behind by only three and could tie with a field goal. Had they missed, they would be behind by five points and still need the same touchdown.

When  $p_1 < 2p_2$ , the optimal strategy attempts to give a team an opportunity to win in regulation time. The two-point conversion can become an integral part of the offense. This can best be seen by comparing the "behind by 1" lines of Tables 1 and 2.

The data of the last three years suggest that under the new rule, it may well be that  $p_1 < 2p_2$ . It is certainly likely to be true for some teams. The most dramatic impact of the new rule can easily occur near the very end of a game. A number of factors come into play that did not have to be considered before. An example is the common occurrence of a team being behind by seven points toward the end of a game.

**Table 2—Optimal Strategy Chart for Seasons 2015 and 2016**

Behind by	Total Possessions Remaining in the Game																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
10			2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
9			2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
8			2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
7			1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2
6			1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
5			2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
4			1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
3			1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Ahead by																					
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
1		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
4		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
5		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
6		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2
7		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
8		1	1	1	1	1	1	1	1	2	1	2	2	2	2	2	2	2	2	2	2
9				1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
10				2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

The optimal strategy chart when FG and TD probabilities are: 0.13 and 0.21, respectively,  $p_1 = 0.938$  and  $p_2 = 0.495$ . 1 means try for one point; 2 means try for two points; blank means it doesn't matter.

Suppose that team gets possession of the ball for one final drive and manages to score a touchdown with “little time” remaining. Standard procedure has always been to try to send the game into overtime by kicking an extra point. What if it were true that  $p_1 < 2p_2$ ? Suddenly there are real issues to consider.

It becomes important to know how much “little time” is left and what can be done in that time. If regulation time runs out as the

touchdown is scored, then the reasoning is simple: Going into overtime is worth only 1/2 of a win. By attempting a one-point conversion, that team would have a  $p_1$  chance of obtaining a 1/2 chance of winning the game. That is, they have a  $\frac{1}{2} p_1$  probability of winning. If they attempt a two-point conversion, there would be no overtime and their probability of winning would be  $p_2$ , which is greater than  $\frac{1}{2} p_1$ .

If time has not completely run out, there is a subtle issue to be considered. After the PAT attempt, when the other team receives the kickoff, they will either be tied, ahead by one, or behind by one. With “little time” remaining, teams behave very differently in the three situations.

- If the other team is behind and believes it to be their last possession, they will be

in desperation mode. They will behave differently than if the game were tied. They will always use all four of their downs to try to advance the ball. They will never punt. They will be more willing to try for a long, low-probability field goal.

- If the score is tied, they may even play more conservatively than usual.
- If they are ahead, they will simply try to run out the clock.

The upshot is that the likelihood that they score points is greater if they are behind than if the score is tied. This complicates the PAT decision since if the other team scores, the game is lost.

If one knew how the scoring rates would change, a simple calculation could resolve the issue. Even bounds on the changed scoring rates would help. These depend heavily on factors such as the exact time remaining and the number of time-outs the other team has. Fortunately, a team having to operate in any of these three modes at the end of a game is not uncommon—these situations have all occurred many times since the beginning of the NFL independently of which PAT rule was used.

In this situation, the data on which win probabilities are based should be helpful. Input based on a coach's experience would be relevant and crucial to any calculations one would want to do.

## Final Thoughts

Suppose the NFL, or just a few teams, were eventually to acknowledge the possibility that  $p_1 < 2p_2$  and that there are many, many more situations when going for two points is appropriate; furthermore, that many of those occur in the first three quarters. How in the world are they ever going to identify those situations without some math?

To do it empirically, the entire league would probably have to commit to 10 years of data collection while making all their PAT decisions by flipping a coin.

It is true that there are, in total, very few games per year in which a non-obvious PAT decision will actually have an impact on the result. In fact, an individual team is not likely to experience the opportunity more than once every few years.

However, the NFL season consists of only 16 regular games. Based on their outcomes, 12 of the 32 teams make the playoffs and retain hopes of reaching the Super Bowl. A coach's job security often depends on his record of making the playoffs. It is very competitive, and it is common for one extra win or loss to determine a team's fate.

NFL teams have the reputation of looking for any way of giving themselves even the least bit of advantage. I continue to find the lack of curiosity about a probabilistic approach astonishing. Yet "the chart" maintains credibility. Little matter that no one really knows how it was developed, what it is based on, or at what point in the game it is appropriate for use. And, of course, it was developed under a different set of rules. ■

## About the Author

**Harold Sackrowitz** is a Distinguished Professor of Statistics at Rutgers University and a Fellow of the ASA and IMS. His current research interests include decision theory and multiple testing.

# An Analysis of the First Round of the MLB First-Year Player Draft

Gabriel Chandler and Simon Rosenbaum

The First-Year Player Draft is a yearly event, held near the beginning of each June. It is the primary mechanism for amateur (primarily high school and college) players to be assigned to the 30 Major League Baseball (MLB) team organizations. It consists of a total of 40 rounds, with draft position being “reverse order”—i.e., teams with worse records from the previous year receive higher placement within the rounds than those with better records.

This is intended to promote competitive balance within the league. Drafted players who opt to turn professional receive a signing bonus, the value of which correlates highly with their draft position, and—nearly always—are assigned to minor league affiliates of the club that drafted them.

After the 2012 season and draft, the MLB implemented a new system governing free agent signings. Once a player becomes eligible for free agency, his team may offer him a “qualifying offer,” which is a single-year contract equal to the average of the top 125 salaries from the previous season.

For the 2017 season, the qualifying offer was approximately \$17 million. Should the player decline the qualifying offer and instead sign with another team, the signing team is required to give up a draft pick. This would be their first-round pick, unless that pick should fall among the first 10 selections. Such picks are known as *protected*.

In the case of a protected pick, the signing team will instead lose their second-round selection. The team from whom the qualifying offer was declined would be compensated with a “sandwich pick” (a selection coming between the first and second rounds of the draft).

It is well understood that higher-round picks are more valuable to a team, since the pool of available talent becomes more shallow as players are selected. For example, MVPs of both the 2015 and 2016 seasons were first-round picks (Josh Donaldson, the 2015 American League MVP, was a supplementary first-round pick; i.e., a sandwich pick). However, the language of the rule privileges the first 10 picks and suggests that there is substantial variability within even the top 30 selections.

The primary goal of this article is to understand how much variability there is within just the first round of the draft, and whether there was justification for protecting the first 10 picks (or whether 10 is just a convenient round number). A simple look at the performance of 10th- versus 11th-pick players, described below, yields strong statistical significance ( $p\text{-val} = .003$ ).

The central technique we use for this analysis is isotonic regression. In the case of a “positive” measure of performance, we search over the class of all decreasing functions, since the explanatory variable is draft slot. Linearity does





Isotonic regression, like linear regression, attempts to find a curve that lies “closest” to a given data set, where closest is defined in terms of sums of squared deviations. However, rather than searching within the class of linear functions, here we only require that the resulting curve is monotone (non-decreasing), so our estimate is non-parametric. Note that a non-increasing curve, as used here, can be found by fitting the negative of the response variable.

not seem an appropriate model for this relationship, since we might not make much distinction between a 299th and 300th pick, but would for a first versus second pick. Thus, the class of decreasing functions seems particularly well suited.

No team should favor a later pick over an earlier one with an expectation to do better, since the set of players available at the later pick are nested inside the set at the earlier pick. As teams attempt to pick the best players possible at each step, it seems reasonable that the expected value of a player should, thus, decrease in the draft slot.

We acknowledge that occasionally a team will pass on a player due to concern about their ability to sign him. We suspect this happens relatively rarely and across many draft slots, and thus would show up in our model primarily as added noise.

Two assumptions are obvious and implicit with using a monotonic function to describe the relationship between draft slot and production. The first is that particular teams are not better than others at predicting productivity and the second is that teams are not drafting for need.

Regarding the former, we examine this at the end of the article and find little evidence of such an

effect. Combined with the fact that draft position for a team changes significantly from year to year, any differences would manifest themselves primarily as noise in our analysis rather than signal. We discuss the issue of drafting for need in the analysis section.

To our knowledge, this is a novel application of order-restricted inference in the analysis of sports data, although it seems natural in many settings here. For instance, player value might be thought of as unimodal in age (monotone on either side of the mode). In addition to an estimate of the expected performance as a function of draft slot, we also look at the significance of differences in the regression function.

## Data

We use as our metric of performance Wins Above Replacement (WAR, specifically Sean Smith's rWAR, publicly available at *baseball-reference.com*), which attempts to summarize the total contribution of a player through all aspects of the game, and allows for the comparison of pitchers with position players. A replacement player is defined as someone a team could acquire easily and for minimal cost. WAR then attempts to measure how many additional games a team won due to having this player rather than a replacement.

Draft data were collected from the June draft each year from 1976 to 2000. Although data are available beyond 2000, we stopped here because later drafts have players for whom career WAR is still very much in flux. We consider the first  $k = 50$  picks from each of those years. Entries for players who never reached the major leagues are missing. This is true for 491 of the 1,250 entries ( $\approx 39\%$ ), giving an indication of how difficult it is

to predict future performance of amateur baseball players.

Figure 1 shows how this proportion varies from year to year. There is some indication that the addition of the Mariners and Blue Jays in 1977 caused the success rate to increase (as one would expect with the need for an additional 50 major leaguers), but no such effect is apparent for the 1993 or 1998 expansion. This may be due to the increase in international players (non-USA/Canada), who are not eligible for June draft, but have become increasingly common on major league rosters (currently more than 25%).

There is a clear need for data imputation for the researcher to be able to handle the missing data values. While a value of 0 would seem sensible, nearly a third of players (251 of 759) who reached the major leagues had WAR values that were negative. It seems reasonable that a player who reached the major leagues was at least as valuable as a player who did not. Thus, we assigned all draftees who failed to reach the majors a value of -4, the smallest observed record in the data set (obtained by pitcher Brian Williams, the 31st pick in the 1990 draft). We acknowledge that our choice is debatable, with other choices likely to yield somewhat different results.

## Analysis

Before fitting a monotonic function to the data, we attempt to allay fears regarding possible non-monotonicity in the relationship, due to teams drafting according to need. Even if teams had the same information available and used it efficiently, there is concern that teams might draft a player who is not the top player available because of positional need (or lack thereof). We examine this by considering how likely a team is

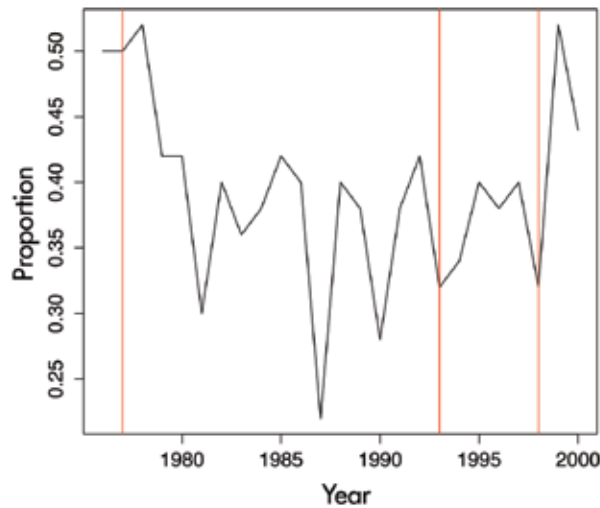


Figure 1. Proportion of drafted players who failed to reach the major leagues, by year. Vertical lines correspond to expansion years (two teams added each occasion).

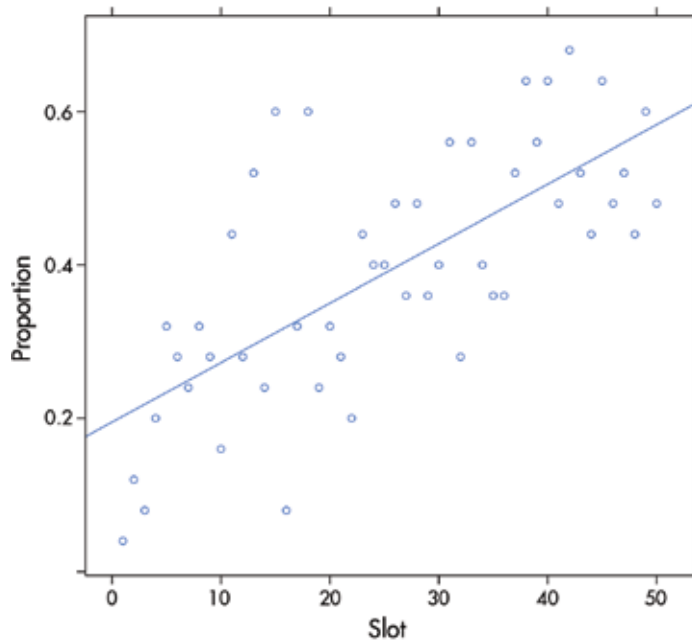


Figure 2. Proportion of drafted players who failed to reach the major leagues, by draft slot, with linear least squares fit.

to draft a pitcher given their team ERA (which indicates how much they need pitching help).

We specifically model the probability a team drafts a pitcher with their first pick (since they may have multiple picks in the first round) against their team ERA. With a sample size of  $n = 985$ , a logistic regression yields a  $\hat{\beta}_1 = 0.157$  ( $p < 0.01$ ) which corresponds to an odds ratio of 117%, meaning

that a team with an ERA a full point higher (a difference of 1.78 standard deviations) only has a 17 percent greater chance of drafting a pitcher. If one looks only in the last decade,  $\hat{\beta} = 0.059$ , which is highly non-significant. We conjecture this is at least partly due to the extended minor league grooming of almost all players.

We also note that ERA has a very weak correlation with draft

position, again meaning that any effect on our analysis will be seen primarily as added noise and less power of subsequent hypothesis tests.

The natural expectation is that higher-drafted players would have a higher likelihood of a productive major league career than players drafted later. Figure 2 shows the proportion of each draft slot in not reaching the major leagues.

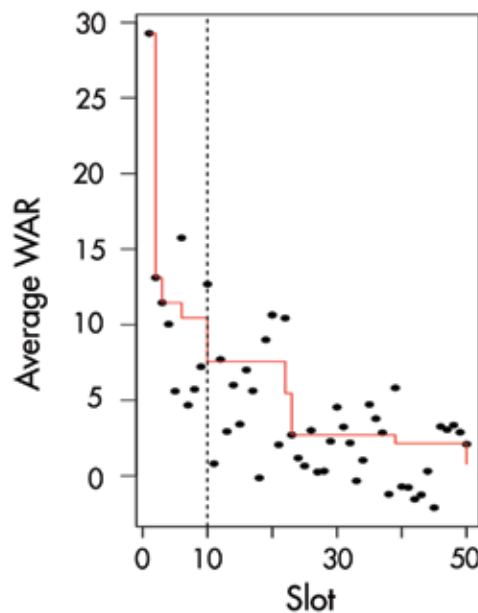


Figure 3. Average (over years) WAR versus draft slot and resulting isotonic regression curve. Dashed vertical line at slot 10, the largest protected pick. Note that the function is continuous from the left.

A linear fit exhibits a lack of fit for very early picks, especially the first three overall selections. One may recognize the names Steve Chilcott and Brien Taylor, the only two first overall picks (1966 and 1991 respectively) not to reach the major leagues.

Of course, drafting a player who simply reaches the major leagues should not be the goal of a team in the first round, so the preceding figure does not capture the whole story. With signing bonuses in the millions of dollars for first-round draft choices, the expectation is for a productive major leaguer. Thus, we seek to understand the relationship between draft slot and expected WAR.

As argued in the introduction, this relationship is clearly monotonically decreasing, due to the set of players available for draft at any slot being a decreasing nested collection of sets. We seek to find the best (in the sense of least squares) monotonically decreasing function

for our data. That is, we consider the WAR at pick  $i$  to be of the form  $Y_{ij} = \mu_i + \epsilon_{ij}$ , with the constraint that  $\mu_i \geq \mu_{i+1}$  for all  $i$ . In other words, the expected WAR decreases as the draft slot increases.

As in linear regression, we seek the solution that minimizes the sum of squared residuals. Wonderfully, it turns out that this solution can be found rather simply, despite the need to search over a much bigger class of functions than linear regression. The resulting estimated regression function is presented in Figure 3. We see the regression function make a large jump at pick 10, which is the last protected pick under the current MLB rules.

One may conjecture that this feature in the data led to the rule. This jump is largely due to the poor average performance of 11th picks (the worst of any pick among the first 17) and the surprising success of 10th picks (higher than six of the preceding nine slots). One should ask how much of the

observed effect is due to chance. Is there strong statistical evidence to suggest that something interesting is happening between the 10th and 11th picks?

As a simple analysis, we run a two sample t-test with the alternative hypothesis of  $H_1: \mu_{10} > \mu_{11}$ , which returns a  $p$ -value of .003 and a 95% lower confidence bound of 4.95 (note the severe lack of normality in the data, however). Whether an increase in career WAR of 4.95 is meaningful is an open question, as well as whether a more-suitable definition of protected picks may exist. One might also wonder whether this result is a consequence of “data snooping.”

To this end, we wish to construct hypothesis tests for decreasing (versus a null of constancy) value over subsets of picks.

Consider picks  $c$  to  $d$ . The null hypothesis states that the expected WAR is constant over these picks. In other words, after removing the top  $c - 1$  prospects from the draft,

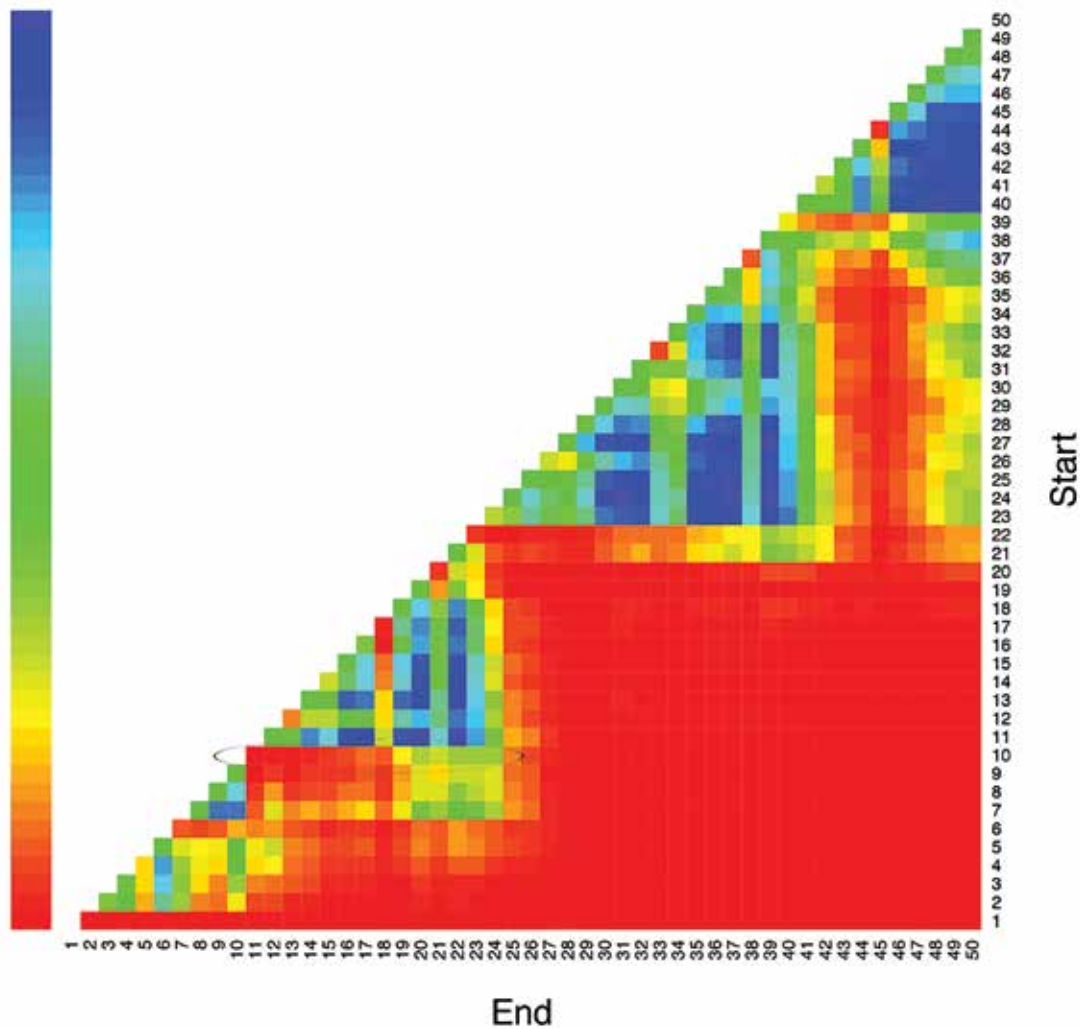


Figure 4. Heat map of  $p$ -values for test of constancy of response curve over different ranges of draft slots. Scale goes from 0 (red) to 1 (blue). Tests comparing 10th pick to later picks are circled.

there are at least  $d - c + 1$  players who are indistinguishable in terms of their “projectability” and agreed to be the best prospects remaining in the draft class.

In such a scenario, one can view these picks as teams essentially randomly selecting one of these players. We can then simulate from this, by simply taking all players drafted between picks  $c$  and  $d$  (inclusive) and permuting their draft slot. We then use isotonic regression to estimate the response curve, and see how far it falls from the true response curve

(which under the null is constant). Similarly, we can compute this distance for the unpermuted data, and compare the two. The proportion of time the permuted data results in a distance greater than the original data is an estimate of the  $p$ -value.

Figure 4 displays the results of our permutation scheme on the data set, with  $m = 1,000$  permutations done for each test. We can see that there is evidence of a difference between the 10th and 11th picks ( $p = 0.002$ ), but this seeming difference is short-lived. There is

not significant evidence to claim that the response curve is decreasing on the set  $[10, t]$  for  $t = 16, \dots, 26$ , even without accounting for multiple testing.

As we might have expected from the discussion thus far, the first pick is special (although interestingly Ken Griffey, Jr. was the first Hall of Fame inductee who was drafted first overall), since the expected WAR is statistically significantly higher than any other pick.

Another interesting observation is the large block of red in the

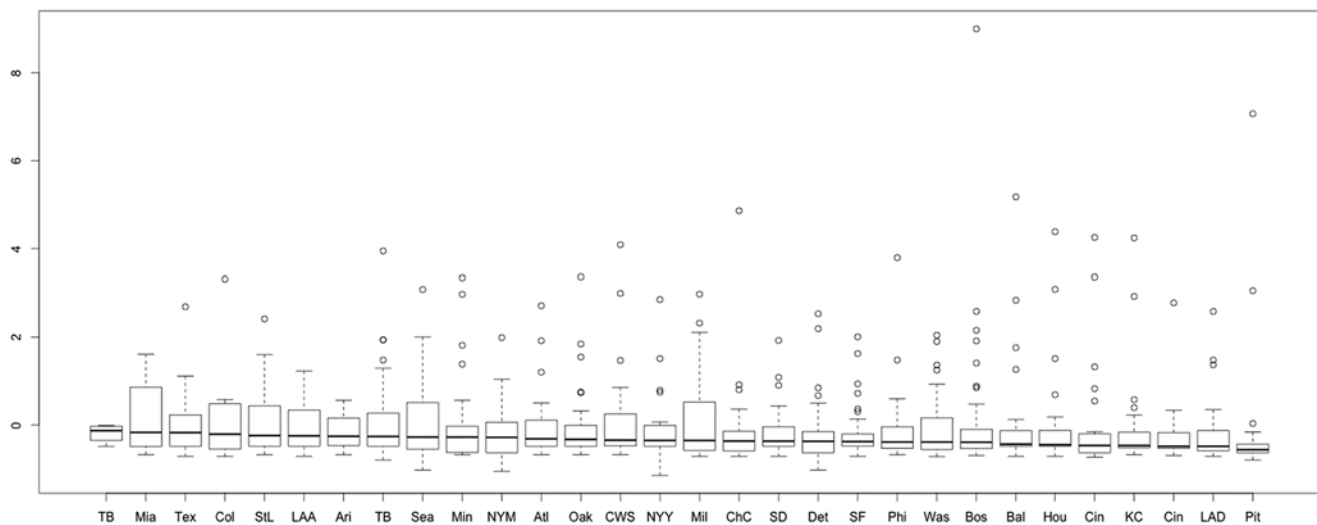


Figure 5. Box plot of z-scores for each player in data set, ordered by 75th percentile. Note: Tampa Bay and Arizona had only four picks each in the data set.

lower right-hand corner of Figure 4. The block covers starting picks from about 1 through 20, and ending picks between 27 and 50. This suggests that any pick in the top 20 should be preferred over any pick 27 or later. Thus, if we were to subscribe to the idea that some picks should be protected regarding free agent signings, a choice somewhere in the range of 20 seems empirically justified.

While the top two thirds of first-round selections may seem like a lot, it would protect the first round pick of exactly the teams to not make the playoffs the previous year (including wild cards, 10 out of 30 teams make the playoffs). This would remove the deterrent of a forfeited first-round draft pick for all non-playoff teams, which would seem to make them more likely to pursue these types of free agents. In debating whether to sacrifice the present for the future, or vice-versa, this sacrifice would be lessened under this definition for such teams.

With these teams improving, the competitive balance of the MLB would increase, which is one of the goals of a reverse-order draft.

Interestingly, the most-recent version of the collective bargaining agreement, announced in late 2016, does away with the protected picks of 1 through 10. Rather, signing teams will now lose either a second- and fifth-round pick, or just a third-round pick, depending on their total team salaries.

## Conclusion

We have explored the value of draft position in the case of the MLB June draft. In particular, we were interested in the justification for the definition of “protected” picks being the first 10 picks in the draft. We conjecture that this value came about in one of two ways.

The first is simply acknowledging that earlier picks are more valuable (and this is certainly true for the first pick overall), and then choosing 10 because it is a “nice” number. The second is that

some empirical study was actually conducted, but may have been short-sighted in scope. Historically, 10th picks have done better than their counterparts one pick later, and the result is statistically significant. However, this does not hold once later picks have been incorporated, since there is no significant evidence that the expected WAR is decreasing on picks 10 through 26.


Had statistical significance been present, the obvious follow-up question would have been whether the difference in WAR was also a meaningful difference. While avoiding this question, we instead advocate for protected picks to include the first 20. This both has a better empirical justification and benefits only non-playoff teams, which promotes competitiveness.

Acknowledging that there is a difference within even the first 50 picks of the draft, we might wonder if there is evidence that one team is doing a better job with their high picks than another. Without this observation, we may



conclude that the Seattle Mariners are draft geniuses with Ken Griffey, Jr. (1987, 83.6 career WAR) and Alex Rodriguez (1993, 118.9 career WAR up to end of 2015 season). These selections seem less impressive, however, when we account for the fact that they had the first overall pick for both of these years. To examine this, we computed z-scores for each player in the data set, and grouped them by drafting team.

The standard deviation, assumed to be changing by pick, was also estimated via an isotonic regression (we assume the variability is decreasing in pick number, which was empirically confirmed). Figure 5 displays comparative box plots for each team in terms of their draftees. We note the largest outlier as Roger Clemens, whose 139.4 WAR was very unusual for a 19th pick.

This data set is also in the “pre-Moneyball” era, which suggests that, given modern analytics, the players in our data set may have been drafted in a considerably different order, again altering the results. It will be interesting to compare these results with those of the modern era, once those data become available. Finally, one can reasonably expect somewhat different results should a metric of performance other than career WAR be used. 

## Further Reading

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., and Brunk, H.D. 1972. *Statistical inference under order restrictions. The theory and application of isotonic regression*. London-New York-Sydney: Wiley Series in Probability and Mathematical Statistics.

Burger, J.D., and Walters, S.J.K. 2009. Uncertain prospects: rates of return in the baseball draft, *Journal of Sports Economics* 10, 485–501 (draft). All Theses. Paper 2088. [http://tigerprints.clemson.edu/all\\_theses/2088](http://tigerprints.clemson.edu/all_theses/2088).

Spurr, S.J. 2000. The baseball draft: A study of the ability to find talent, *Journal of Sports Economics* 1, 66–85.

## About the Authors

**Gabe Chandler** is an associate professor in the Department of Mathematics at Pomona College.

**Simon Rosenbaum**, a 2016 graduate of Pomona College in mathematics and economics and a 2014 NCAA Division-III first team All-American baseball player, is an assistant in baseball development for the Tampa Bay Rays.

# An Ordinal Logistic Regression Model for the Masters Golf Tournament

*Erik L. Heiny and Cody C. Frisby*

**I**t is April of 2016 and Jordan Spieth, the defending champion and second-ranked player in the world, is walking toward the 10th tee in the final round of the Masters. Spieth has moved to seven under par coming off four straight birdies to close out his front nine, and has a five-shot lead. His pursuers include Jason Day, the number-one ranked player in the world, but Day is seven shots back at even par and beginning play on number 11.

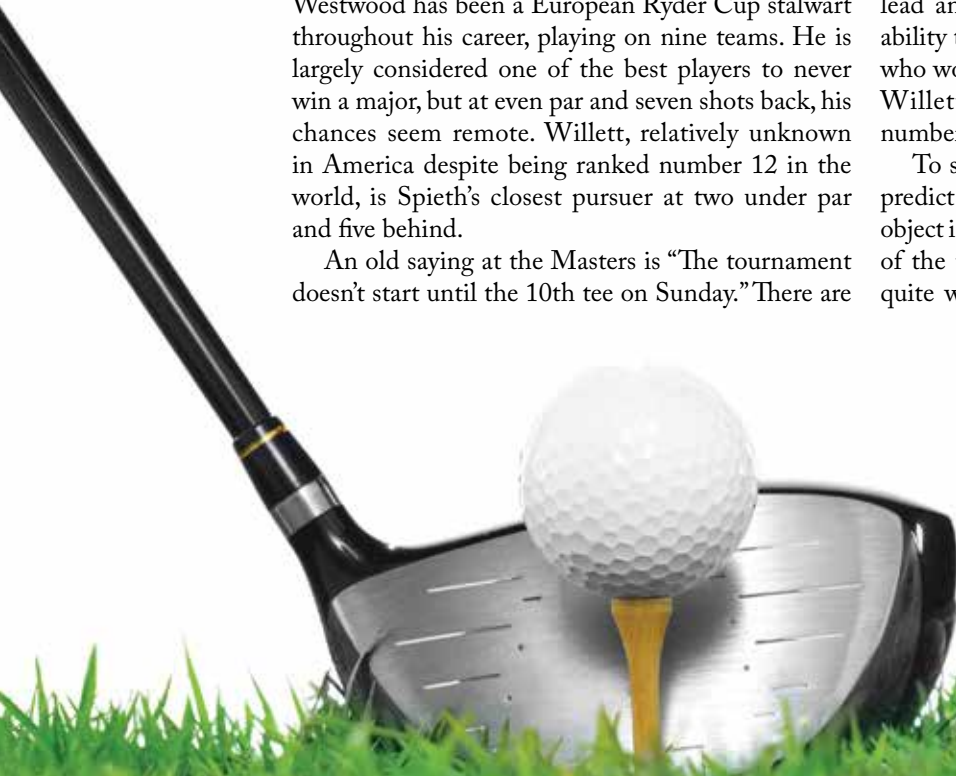
Paired with Day is Dustin Johnson who is six shots back at one under par. On the 12th tee are two Europeans: Lee Westwood and Danny Willett. Westwood has been a European Ryder Cup stalwart throughout his career, playing on nine teams. He is largely considered one of the best players to never win a major, but at even par and seven shots back, his chances seem remote. Willett, relatively unknown in America despite being ranked number 12 in the world, is Spieth's closest pursuer at two under par and five behind.

An old saying at the Masters is "The tournament doesn't start until the 10th tee on Sunday." There are

eagle and birdie opportunities on the back nine, but also water hazards and potential big numbers. In fact, the cut at the Masters is to the low 50s and ties, but due to how quickly players can make up or lose ground at Augusta National, anyone within 10 shots of the lead after two rounds also makes the cut (even if outside the top 50).

But we are talking about Jordan Spieth here. He knows how to win at Augusta, he won the Masters and U.S. Open the year before, and was in contention at the British Open, the third leg of a potential grand slam, all the way to the 72nd hole. With a five-shot lead and only nine holes to go, what is the probability that Spieth doesn't win? If Spieth doesn't win, who would be most likely to catch him? Would it be Willett, his closest pursuer, or maybe Day, the number-one player in the world?

To simulate the Masters, it might be tempting to predict round-by-round scores for each player. If the object is simply to predict who will win before the start of the tournament, this approach may indeed work quite well. However, it cannot answer the question



posed above. In a 2010 article, McHale assessed the fairness of the golf handicapping system in the UK by using ordered logistic regression to model hole-by-hole scores as a function of covariates. He then used this model to simulate scores for players of different handicaps and estimate probabilities of winning hole-by-hole matches for each player.

This article follows McHale's lead and use ordinal logistic regression to produce probabilities of eagle, birdie, par, bogey, double bogey, and triple bogey, for each player, on each hole, in each round, for every year. Simulations of the Masters tournament can then be conducted to answer several interesting questions. How well does this model forecast the Masters? What are the best predictors of Masters success? Is it driving or putting, or just world ranking? How has Tiger Woods's probability of winning changed over the years?

In his prime, a common question was whether you should take Woods or the rest of the field? Was Woods's probability of winning really high enough to make this a reasonable question?

Leading into the 2017 Masters, Dustin Johnson was playing the best golf of his life. He had won his last four starts and was the number-one ranked player in the world. Then he had an accident the day before the Masters, injured his back, and had to withdraw. What were his chances of winning?

Finally, what were Spieth's chances of winning back in 2016? Ordinal logistic regression and simulation will help answer all these questions.

## Methods

Hole-by-hole scores were available for each Masters participant from 1983 to 2017 using the PGA Tour's ShotLink™ database. The key fields in this data set were player, year, round, hole, par, and score relative to par or RTP score. The other covariates used in the model were strokes gained (SG), which are multiple measures of player skill developed by the PGA Tour; official world golf rankings (OWGR); and number of Masters appearances.

The object of our model is to develop probabilities of different scores for each player on each hole they play. The possible scores on any hole are eagle or better (2 shots less than par or RTP = -2), birdie (1 shot better than par or RTP = -1), par (RTP = 0), bogey (1 shot more than par or RTP = 1), double bogey (2 shots more than par or RTP = 2), and triple bogey or worse (3 shots or more than par or RTP = 3).

Since we want to estimate probabilities for an ordinal response, score, an ordered logistic regression model is appropriate. Ordinal logistic regression takes

advantage of the cumulative nature of the response by using cumulative logits. A common type of cumulative logit, referred to as a proportional odds model (POM), is shown as defined in *Categorical Data Analysis* by Agresti:

$$\ln\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \alpha_j + \beta'x \quad (1)$$

Notice that *cumulative* probabilities are being modeled here.  $Y$  represents the player's score on a hole,  $j$  goes from 1 to  $J-1$  ( $J = 6$  in this study, the number of possible scores on a hole),  $\alpha_j$  is the intercept for the  $j$ th cumulative logit,  $\beta$  is a vector of coefficients for the covariates or fixed effects, and  $x$  is the vector containing values for the covariates which are year, round, hole, number of Masters, strokes gained, and official world golf ranking. These cumulative logits can be easily converted into cumulative probabilities and then into probabilities for each possible score, as demonstrated below.

$$P(Y \leq j) = \frac{1}{1 + \exp(-\alpha_j - \beta'x)} \quad (2)$$

$$P(Y = j) = P(Y \leq j) - P(Y \leq j - 1) \quad (3)$$

$$P(Y = 1) = P(Y \leq 1) \quad (4)$$

$$P(Y = J) = 1 - P(Y \leq J - 1) \quad (5)$$

For this study, it is important that the covariates in  $x$  are good predictors of score. Let's take a closer look at the covariates being used in this model. The actual hole being played has a big impact on score. All golfers know that different holes on a course will vary in difficulty. Figure 1 is a plot of scores by year for hole number 2; a par five where the green is reachable in two shots; and hole number 18, a long uphill par four. It is clear that hole 2 will tend to produce lower scores relative to par than hole 18. To account for the varying difficulty of each hole at Augusta National, hole is included as a categorical variable in the model.

Year and round are also included as categorical variables. The course will vary in difficulty from one year to the next, and from one round to the next. These differences are most likely due to weather conditions, firmness of the greens, pin locations, etc. In the 2017 Masters, there were very high winds during the first two rounds, and then the weekend had almost perfect weather. A 2013 article by Balsdon gives some evidence that PGA Tour players will change risk strategies and playing styles when near the cut line, which would affect round two scores.

Regardless of the reasons, including year and round helps account for a significant portion of the

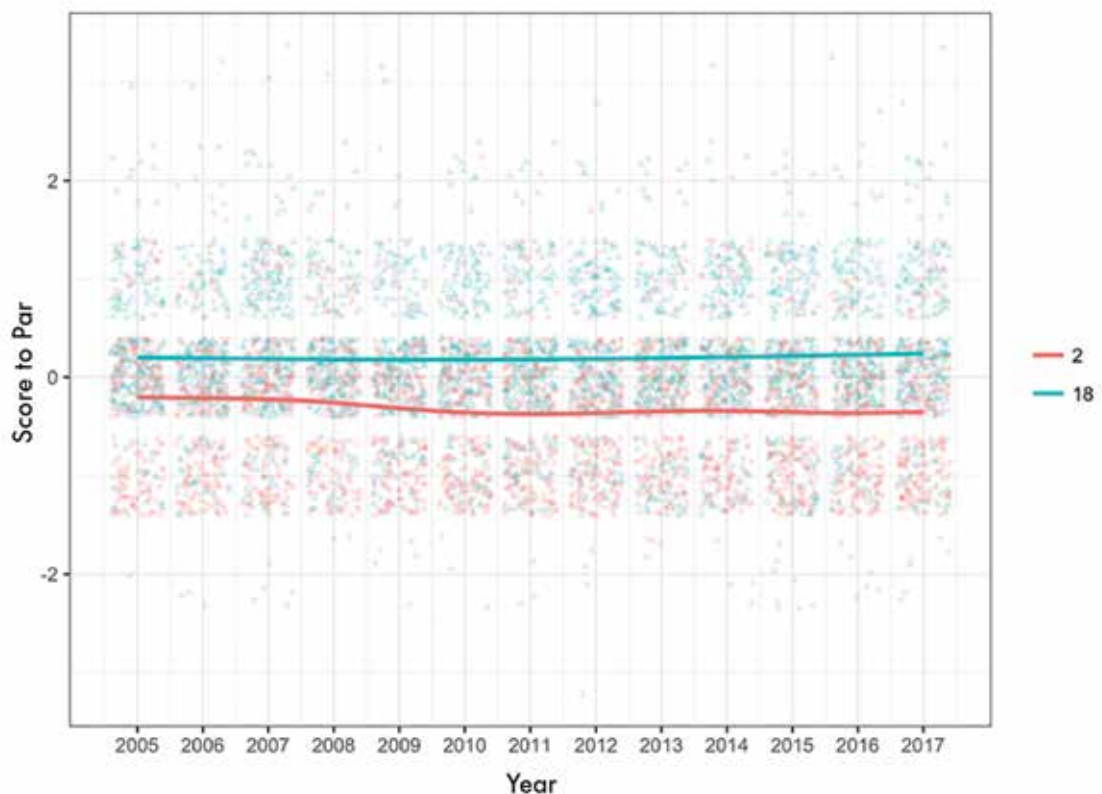


Figure 1. Scoring on holes 2 and 18 by year.

variability in players' scores, and provides a more-precise estimate of how players' specific skills—such as strokes gained, world ranking, and number of Masters appearances—affect score.

The strokes-gained concept, initially developed by Mark Broadie of Columbia University, measures PGA Tour players' skills in different categories relative to the average player on the tour. The different skill categories, as defined on the PGA Tour website ([www.pgatour.com/news/2016/05/31/strokes-gained-defined.html](http://www.pgatour.com/news/2016/05/31/strokes-gained-defined.html)), are:

- *Strokes Gained Off-the-Tee (SG Tee)*: measures player performance off the tee on all par fours and par fives.
- *Strokes Gained Approach-the-Green (SG Approach)*: measures player performance on approach shots. Approach shots include all shots that are *not* from the tee on par four and par five holes and are *not* included in strokes gained around-the-green and strokes gained putting. Approach shots include tee shots on par threes.

- *Strokes Gained Around-the-Green (SG Around)*: measures player performance on any shot within 30 yards of the edge of the green; does *not* include any shots taken on the putting green.
- *Strokes Gained Putting (SG Putt)*: measures how many strokes a player gains (or loses) on the greens.

Note: When fitting the model, SG Tee was set equal to zero on all par threes because it only measures skill off the tee on par four and par five holes.

To illustrate how these statistics work, consider Sergio Garcia, the 2017 Masters champion. His SG Tee value was 0.737 for the 2016 PGA Tour season. That means that over the course of one round, Garcia was 0.737 shots better than the average PGA Tour player with his tee shots on par four and par five holes. For a full tournament with four rounds, Garcia would pick up about three shots off the tee, relative to an average player. In the other strokes gained

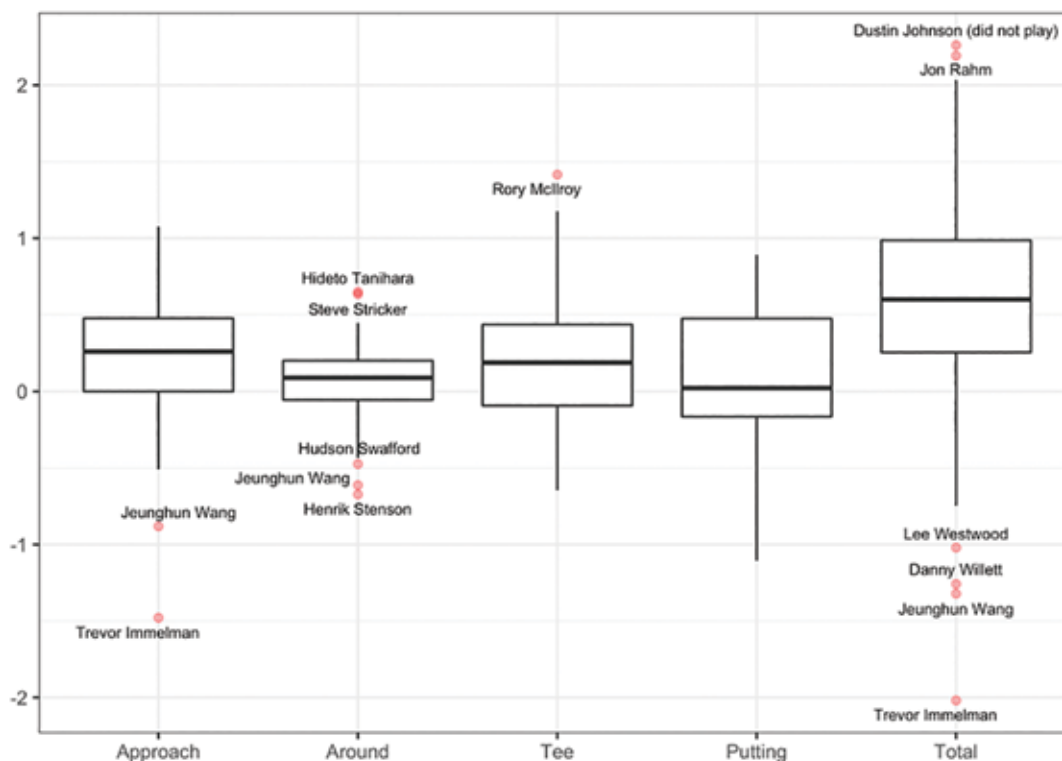


Figure 2. Strokes gained data for 2017 Masters.

categories, Garcia's values were SG Approach = 0.482, SG Around = -0.031, and SG Putting = -0.388. The strokes gained statistics are additive, so Garcia's SG total =  $0.737 + 0.482 - 0.031 - 0.388 = 0.801$ . That means that overall, Garcia was 0.801 strokes better per round than the average player on the PGA Tour.

It should be noted that strokes gained are only computed for selected rounds in PGA Tour events. For Garcia in 2016, his statistics were based on 31 measured rounds. Additionally, the strokes gained statistics in this study were adjusted so they represented the prior 12 months of play leading into the Masters. Since strokes gained data are only available going back to 2004, the Masters tournaments from 2005 to 2017 are included in this study.

Leading into the 2017 Masters, Garcia's strokes gained values over the previous 12 months were adjusted to: SG Tee = 0.957, SG Approach = 0.085, SG Around = 0.148, SG Putt = -0.373, and SG total = 0.817. These values were based on 30 measured rounds.

Some players were removed from the data set due to missing or limited strokes gained data. The Masters

field always includes several older former champions, as well as younger amateurs who do not play on the PGA Tour at all. To be included in the data set, a player had to have at least 10 measured rounds in the 12 months before the Masters. Former champions removed from the 2017 data set include Fred Couples, Bernhard Langer, Larry Mize, Mark O'Meara, Sandy Lyle, and Ian Woosnam. Couples had a high finish in 2017, and Langer was in the second-to-last group for the final round of 2016 before fading, but generally, these players do not have a big impact on the outcome of the tournament.

Figure 2 contains boxplots for all four strokes gained categories, as well as strokes gained total for 2017. The PGA Tour adjusts strokes gained so the average value for each category is zero, but the center tends to be above zero for Masters participants; most notably, strokes gained total. This is not surprising considering the Masters has a limited field and not all PGA Tour players qualify for the tournament.

Something else to note from Figure 2 is how close the skill levels are between these players. Just looking at SG total, the middle 50% are within 0.75 shots of each other per round.



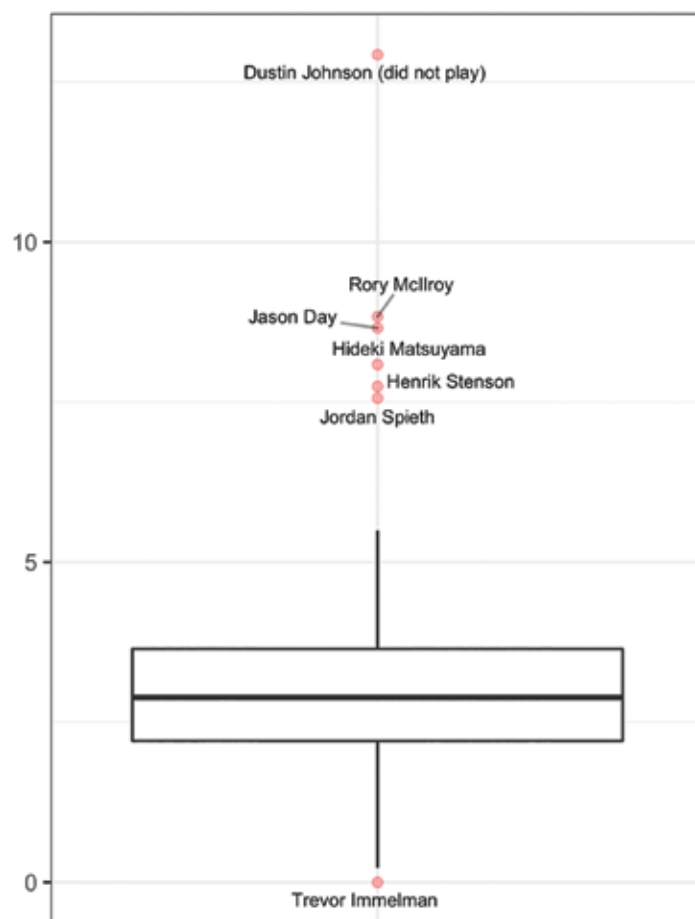


Figure 3. Official world golf ranking of 2017 Masters participants.

Finally, Figure 2 contains the SG values for Dustin Johnson, who did not play in the 2017 Masters due to a back injury suffered the day before the tournament began. Nevertheless, being the number-one player in the world and winner of his four prior starts leading into the Masters, he is included as a player of interest.

The Masters field has always had a significant international presence. Foreign-born champions include Gary Player, Seve Ballesteros, Nick Faldo, Bernhard Langer, Sandy Lyle, Jose Maria Olazabal, and Angel Cabrera, just to name a few. Many of these players do not play full-time on the PGA Tour, and it is important to have a covariate that accounts for play around the globe.

The official world golf rankings are computed each week using a point system that takes into account tournament results from the PGA Tour, European Tour, Asian Tour, Japan Tour, and others. The ranking system is a two-year rolling average that accounts for

field strength, extra points for major championships, and more emphasis on recent tournament results. (For more details, go to [www.owgr.com/about](http://www.owgr.com/about).)

This study used the official world golf ranking (OWGR) points of each player the week of the Masters. Figure 3 is a boxplot showing the distribution for OWGR leading into the 2017 Masters. It is approximately normally distributed, but with the top six players as outliers. As mentioned earlier, Johnson did not play. Similar to strokes gained, the middle 50% are tightly bunched together and within 1.5 points of each other.

The last covariate in the model is the number of Masters played. Rarely do first-time players win the Masters. In fact, it has happened only three times: Horton Smith won the first Masters, played in 1934; Gene Sarazen in 1935 with his “shot heard round the world,” making a double eagle on the par five 15th in the final round; and Fuzzy Zoeller in 1979.

**Table 1—Fit Statistics for Models POM and PPOM**

	POM	PPOM
-2lnL	118,500	117,094
AIC	118,588	117,454
BIC	118,747	118,107
n	58,986	58,986

It usually takes some time to learn how to play Augusta National, to get familiar with the large undulating greens, when to be aggressive, and when to play more conservatively. Including the number of Masters as a categorical variable gave the best results. For a player making a first or second Masters appearance, number of Masters was set equal to 1; otherwise it was 0.

For this data set, there is a clustering of observations, or scores, within each player. To account for this, we use a class of models—generalized linear mixed models (GLMM)—that focuses on inferences about the individuals (or golfers) in the population. These are conditional models (conditional on each player) that allow for estimating probabilities of score on each hole that are player-specific. To do this, we modify equations (1) and (2) above by adding in a random player effect for the  $i$ th player,  $\gamma_i$ , where  $\gamma_i \sim N(0, \sigma_\gamma^2)$ .

$$\ln\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \alpha_j + \beta_j' x + \gamma_i \quad (6)$$

$$P(Y \leq j) = \frac{1}{1 + \exp(-\alpha_j - \beta_j' x - \gamma_i)} \quad (7)$$

It should be noted that observations within each cluster are assumed to be independent. In other words, a golfer's score from one hole to the next is assumed to be independent. Is this reasonable? To answer this gets into the “hot-hand” controversy. Arkes and Clark address this issue on the PGA Tour; Clark found no evidence of correlation in PGA Tour players' scores from one hole to the next, and Arkes found a significant “cold hand effect” in three, six, nine, and 18 hole stretches. However, this effect was very minimal.

For example, Arkes found that for every shot over par in a three-hole stretch, the PGA Tour player would average approximately 0.01 shots over par for the next three holes. Furthermore, empirical evidence for this data set supports the assumption. Autocorrelation at lag one—correlation of score from hole to hole—was equal to 0.0139. In the McHale study regarding the fairness of handicaps in the UK, it was 0.016.

The model defined in equations (6) and (7) is still a proportional odds model. This model enforces the “parallel lines assumptions,” which means that the effects of the  $x$  covariates do not change across the  $J-1$  cumulative logits. If these effects are allowed to change, it is a non-proportional odds model, or a partial proportional odds model (PPOM) if the effects for some, but not all the covariates, change across the  $J-1$  cumulative logits (see equations (8) and (9)).

After testing the “parallel lines assumption” for each effect, the effects for hole, round, year, number of Masters, SG Tee, and SG around were allowed to vary across the  $J-1$  cumulative logits, and the effects for SG Approach, SG Putt, and OWGR continued to use the “parallel lines assumption.”

$$\ln\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \alpha_j + \beta_j' x + \gamma_i \quad (8)$$

$$P(Y \leq j) = \frac{1}{1 + \exp(-\alpha_j - \beta_j' x - \gamma_i)} \quad (9)$$

Both models—POM and PPOM—were fit to the data in this study. The SAS procedures PROC GLIMMIX and PROC NLMIXED were used to fit the POM and PPOM, respectively. As shown in Table 1, the PPOM provided a better fit based on the log-likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). However, due to ease of interpretation regarding covariates using the POM, results from both models are discussed in the following sections.

## Cross-Validation Analysis

To determine the predictive ability of the model, the Masters was forecasted using the PPOM for years 2014, 2015, 2016, and 2017. For 2014, the model coefficients were estimated using data only through 2013. Probabilities of score on each hole were then computed for the 2014 players, and the Masters tournament was simulated 10,000 times using R.

**Table 2—Spearman’s Rho Between Model Rankings, Vegas Rankings, and Actual Finish**

	<b>2017</b>	<b>2016</b>	<b>2015</b>	<b>2014</b>
Model vs. Actual	0.4810	0.3787	0.5290	0.2695
Vegas vs. Actual	0.5250	0.3550	0.5706	0.2267
Model vs. Vegas	0.8887	0.8873	0.8442	0.8928

**Table 3—2017 Masters Predictions for PPOM and Vegas Odds**

<b>Name</b>	<b>PPOM Rank</b>	<b>PPOM Prob.</b>	<b>Vegas Prob.</b>	<b>Vegas Rank</b>	<b>Actual Finish</b>	<b>Score</b>
McIlroy, Rory	1	0.1191	0.1111	2	7	285
Day, Jason	2	0.0864	0.0606	3	22	290
Spieth, Jordan	3	0.0863	0.1333	1	11	287
Matsuyama, Hideki	4	0.0653	0.0435	7	11	287
Rose, Justin	5	0.0483	0.0323	8	2	279
Fowler, Rickie	6	0.0409	0.0526	5	11	287
Scott, Adam	7	0.0336	0.0278	11	9	286
Rahm, Jon	8	0.0331	0.0526	5	27	291
Stenson, Henrik	9	0.0323	0.0303	10	58	150
Kuchar, Matt	10	0.0259	0.0179	16	4	283
Hatton, Tyrrell	11	0.0251	0.0179	16	76	156
Casey, Paul	12	0.0240	0.0244	14	6	284
Mickelson, Phil	13	0.0237	0.0556	4	22	290
Garcia, Sergio	14	0.0235	0.0323	8	1	279

Each simulation had a 36-hole cut to the low 50 and ties, or anyone within 10 shots of the lead, and playoffs were used in case of ties at the end of 72 holes.

The players were ranked according to their estimated probability of winning, and spearman’s rho was computed between the model rankings and the actual finish. For comparison, spearman’s rho was also computed between the Vegas rankings (<http://www.vegasinsider.com>) and actual finish, as well as between the model rankings and Vegas rankings. To make these comparisons the same, the older former champions and amateurs who do not play on the PGA Tour were also removed from the Vegas rankings and actual finish. This process was repeated for years

2015, 2016, and 2017, and the results are included in Table 2.

The model has similar predictive power to the posted Vegas odds before the tournament. In some years, the correlation between the model and actual results is low and in others it is high, essentially depending on how well the favorites played. The strokes gained and world rankings data both show that these players are very close in ability and it is difficult to separate them.

It is important to keep in mind that each Masters played is just one realization of what could have happened that year. Since the model predicts as well as the Vegas posted odds, it gives reasonable probabilities of

**Table 4—Estimated Coefficients for  $\beta$ , the Fixed Effects in the POM**

	<b>Coefficient</b>	<b>p-value</b>
$\alpha_1$ (triple bogey or worse)	-5.4122	< .0001
$\alpha_2$ (double bogey or worse)	-3.3611	< .0001
$\alpha_3$ (bogey or worse)	-0.9032	< .0001
$\alpha_4$ (par or worse)	2.1643	< .0001
$\alpha_5$ (birdie or worse)	5.927	< .0001
Hole 1	0.2138	< .0001
Hole 2	-1.6314	< .0001
Hole 3	-0.6399	< .0001
Hole 4	0.2145	< .0001
Hole 5	-0.05217	0.2888
Hole 6	-0.2388	< .0001
Hole 7	-0.08299	0.096
Hole 8	-1.4304	< .0001
Hole 9	-0.3127	< .0001
Hole 10	0.04909	0.3188
Hole 11	0.4085	< .0001
Hole 12	-0.1878	0.0002
Hole 13	-1.7337	< .0001
Hole 14	-0.3093	< .0001
Hole 15	-1.6608	< .0001
Hole 16	-0.3847	< .0001
Hole 17	-0.1031	0.0368
Hole 18	0	.

*continued on p. 52*

what could have happened for each Masters, and these probabilities are the real interest. Table 3 lists the top 14 players for 2017 to include the actual champion, Sergio Garcia. The table includes rank, estimated probability of winning for both the PPOM and Vegas, and the actual finish and score for each player.

Not included in Table 3 is the probability of winning for world number one, Dustin Johnson. He had a 0.2028 chance of winning when included in the model, and Vegas had him listed at 5 to 1 or a 0.1667 chance of winning.

## Fixed Effects

The cumulative logits in (6) and (8) were designed to find the probability of a triple bogey or worse ( $RTP = 3$  or  $j = 1$ ), then the probability of a double bogey or

worse ( $RTP = 2$  or  $j = 2$ ), etc., up to the probability of a birdie or worse ( $RTP = -1$  or  $j = 5$ ). Of course, the probability of an eagle ( $RTP = -2$  or  $j = 6$ ) is 1 minus the probability of birdie or worse.

The full model was developed using 58,696 observations from years 2005 through 2017. While the PPOM provided a better fit, the fitted POM coefficients are given in Table 4 due to their ease of interpretation relative to the PPOM.

In general, when looking at Table 4, a negative coefficient means that scores will tend to get better (or lower) as the covariate increases, and a positive coefficient means that scores will get worse (or higher). For the continuous covariates—strokes gained and world ranking—all coefficients are negative, which means players of higher skill will tend to have lower

**Table 4—Estimated Coefficients for  $\beta$ ,  
the Fixed Effects in the POM** (continued from p. 51)

	<b>Coefficient</b>	<b>p-value</b>
SG_Approach	-0.0502	0.0437
SG_Around	-0.03668	0.3287
SG_Tee	-0.122	< .0001
SG_Putt	-0.02025	0.4
OWGR_april	-0.02652	< .0001
Round 1	0.139	< .0001
Round 2	0.1813	< .0001
Round 3	0.1403	< .0001
Round 4	0	.
no_masters_category 1	0.04489	0.0245
no_masters_category 0	0	.
Year 2005	-0.04007	0.3488
Year 2006	-0.0423	0.327
Year 2007	0.2997	< .0001
Year 2008	-0.02854	0.5091
Year 2009	-0.2398	< .0001
Year 2010	-0.1284	0.0026
Year 2011	-0.2497	< .0001
Year 2012	-0.08938	0.0329
Year 2013	-0.09805	0.0205
Year 2014	-0.00216	0.9594
Year 2015	-0.2577	< .0001
Year 2016	0.08403	0.0434
Year 2017	0	.

scores. Based on the magnitude of the coefficients and  $p$ -values, SG Tee is the most important of the SG skills at the Masters.

Interestingly, SG Around, and SG Putt are not statistically significant. However, it would be a mistake to conclude that chipping and putting are not important at the Masters. The players who do well each year almost certainly performed well in these areas. It just means that using the chipping and putting performance of a PGA Tour player to predict scoring at the Masters will not be as effective as using their driving performance.

For the categorical covariates, year, round, hole, and number of masters, all coefficients are relative to a

baseline category. The baselines are year equal to 2017, round equal to four, hole equal to 18, and number of Masters more than two. The positive coefficient on number of Masters in Table 4 shows that the inexperienced players were at a disadvantage, and tended to shoot higher scores as expected.

The impacts for hole, round, and year can be more clearly seen in Figure 4. Based on the magnitude of the coefficients, hole had the largest impact on scoring. More specifically, the large negative coefficients for the par fives—holes 2, 8, 13, and 15—indicate they were the easiest holes, and that the players had to do their scoring on the par fives at Augusta. The large positive coefficient for number 11—the long par



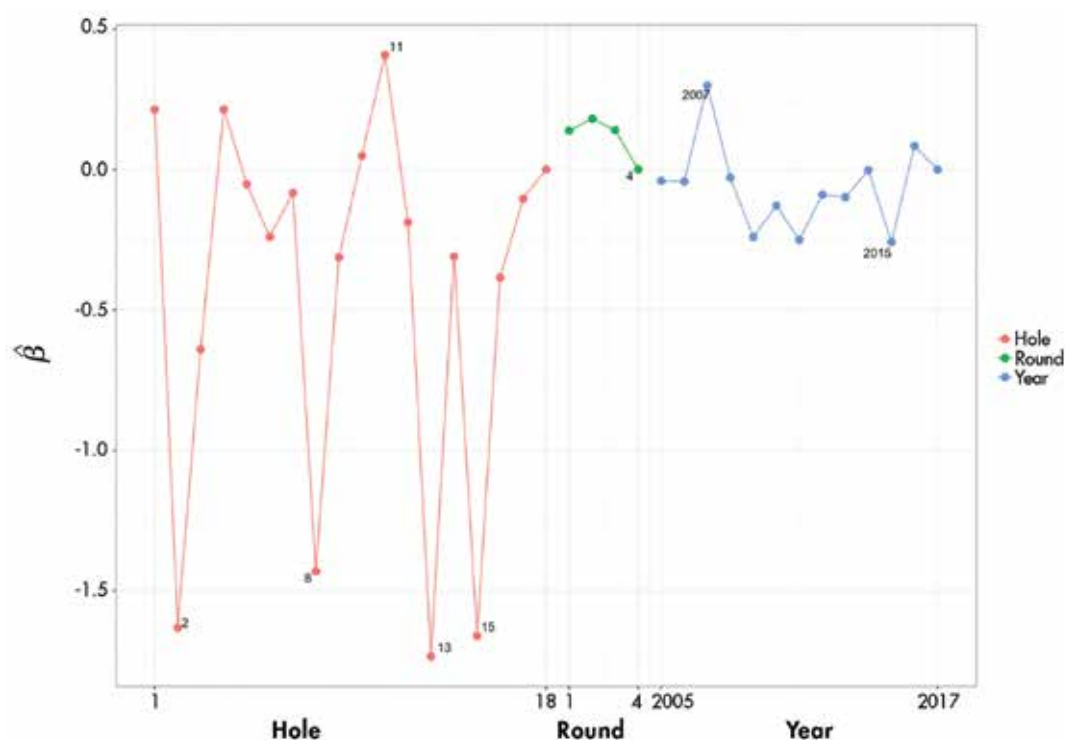


Figure 4. POM coefficients for hole, round, and year.

**Table 5—POM Cumulative Probabilities for Hole 2 vs. Hole 18**

	Hole 2 POM	Hole 18 POM
P(triple or worse)	0.000081	0.004104
P(double or worse)	0.00623	0.031049
P(bogey or worse)	0.068233	0.272344
P(par or worse)	0.611433	0.889411
P(birdie or worse)	0.985458	0.997121

four over 500 yards with a pond front and left of the green—indicates it was the toughest hole.

Year, while not as important as hole, also had a significant impact on scoring. Figure 4 also shows that 2007 was the year the course played the most difficult, and year 2015 was the year it played the easiest. There was cold, windy weather in 2007, and Zach Johnson won with a score of one over par, which was the only year in this study where the winning score was over par. In 2015, Jordan Spieth won at 18 under par, tying Tiger Woods's record score from 1997.

For the round effect, model coefficients indicate that round four had lower scoring than the other three rounds, but this effect was small relative to the effects of year, and especially hole. Any effect due to round is probably a result of course setup; most specifically,

pin placements. Viewers of the Masters will know that announcers often mention “traditional Sunday pin placements” for certain holes, and perhaps tournament organizers want better scoring on Sundays, or at least better chances of eagles for the “Sunday roars.” Regardless, scoring was better in round four over the course of this study.

To see how these coefficients turn into probabilities a little more clearly, consider the difference in holes 2 and 18 for a player playing in 2017's round four with SG equal to zero for all skills, number of masters played is more than two, and OWGR is equal to three (approximately the 50th percentile). Equation 10 below shows the explicit calculation for the first cell in Table 5.

**Table 6—POM Probabilities of RTP score for Hole 2 vs. Hole 18**

	Hole 2 POM	Hole 18 POM
P(triple or worse)	= <b>0.000081</b>	= <b>0.004104</b>
P(double)	0.00623 – 0.000081 = <b>0.006149</b>	0.031049 – 0.004104 = <b>0.026945</b>
P(bogey)	0.068233 – 0.00623 = <b>0.062003</b>	0.272344 – 0.031049 = <b>0.241295</b>
P(par)	0.611433 – .068233 = <b>0.5432</b>	0.889411 – .272344 = <b>0.617067</b>
P(birdie)	.985458 – 0.611433 = <b>0.374025</b>	0.997121 – 0.889411 = <b>0.10771</b>
P(eagle)	1 – .985458 = <b>0.01454</b>	1 – 0.997121 = <b>0.002879</b>

These cumulative probabilities are then easily converted into probabilities for each score.

$$P(\text{triple or worse} - \text{hole 2}) = \frac{1}{1 + \exp(-(-5.4122 - 1.6314 - 0.02652 \cdot 3))} = 0.000081 \quad (10)$$

Table 6 clearly shows that hole 2 is going to yield lower scores relative to par. The reader can easily verify that this player's average score would be 4.194 for hole 18 (0.194 over par) and 4.672 for hole 2 (0.328 under par). These results are consistent with Figure 1.

### Random Effects (Subject-Specific or Player Effects)

In addition to the fixed effects in the model discussed above, random player effects were also estimated. Each player has his own set of values for the player-specific covariates, strokes gained, world ranking, and number of Masters. These alone give each player different score probabilities on each hole. Therefore, in this context, the random player effect, estimated from the clustering of observations within each player, can be interpreted as whether the player plays the Masters better or worse than expected, after adjusting for their skill (SG and OWGR) and experience (no. of Masters).

Figure 5 shows a histogram of the estimated random player effects. For this figure, the sign of the coefficient was changed so positive coefficients indicate that the player tends to have lower scores than expected at the Masters, and negative coefficients indicate the player tends to have higher scores than expected. The outliers in Figure 5—players who play the Masters particularly well—are Jordan Spieth, Justin Rose, Angel Cabrera, and Phil Mickelson.

Spieth has the highest coefficient at 0.0901, and based on his brief record at the Masters to this point, it

is no surprise. Spieth has one victory, two seconds, and an 11th-place finish. Even after adjusting for his high skill level, he plays the Masters better than anyone.

Rose is next at 0.0870. While he does not have any Masters wins, he has finished second twice in the last three years, has five top 10 finishes, and has never missed the cut.

Mickelson is fourth at 0.0830. Even though he has a better Masters record than Rose, he has also had better strokes gained statistics and usually a higher world ranking through the years.

Players of note who performed particularly poorly in the Masters between 2005 and 2017 are Martin Kaymer and Ernie Els. Kaymer, at -0.0496, is a two-time major champion who has missed the cut five times in 10 tries, and just recently had his best-ever finish at 16th in 2017. Els, at -0.0555, is a two-time winner of both the U.S. Open and the British Open. While he placed second in the Masters in both 2000 and 2004, he missed the cut five times between 2005 and 2017, and his best finish is 13th.

### Simulations

For each year from 2005 to 2017, coefficients from the PPOM (for both fixed and random effects) were used to compute probabilities of different scores on each hole and each round for each player. The Masters tournament for each year was then simulated 10,000 times. Again, all simulations involved a cut to the low 50 and ties, or anyone within 10 shots of the lead, and playoffs were run in case of a tie for first place.

From these simulations, probabilities of winning for each player were estimated. These estimated probabilities of winning were used to further investigate some of the covariates in the model. Figures 6 and 7 show the probability of winning versus OWGR and the probability of winning versus SG Total (the sum of all four strokes gained categories), respectively.

These figures have also labeled some players of note, and the Masters champion each year is highlighted

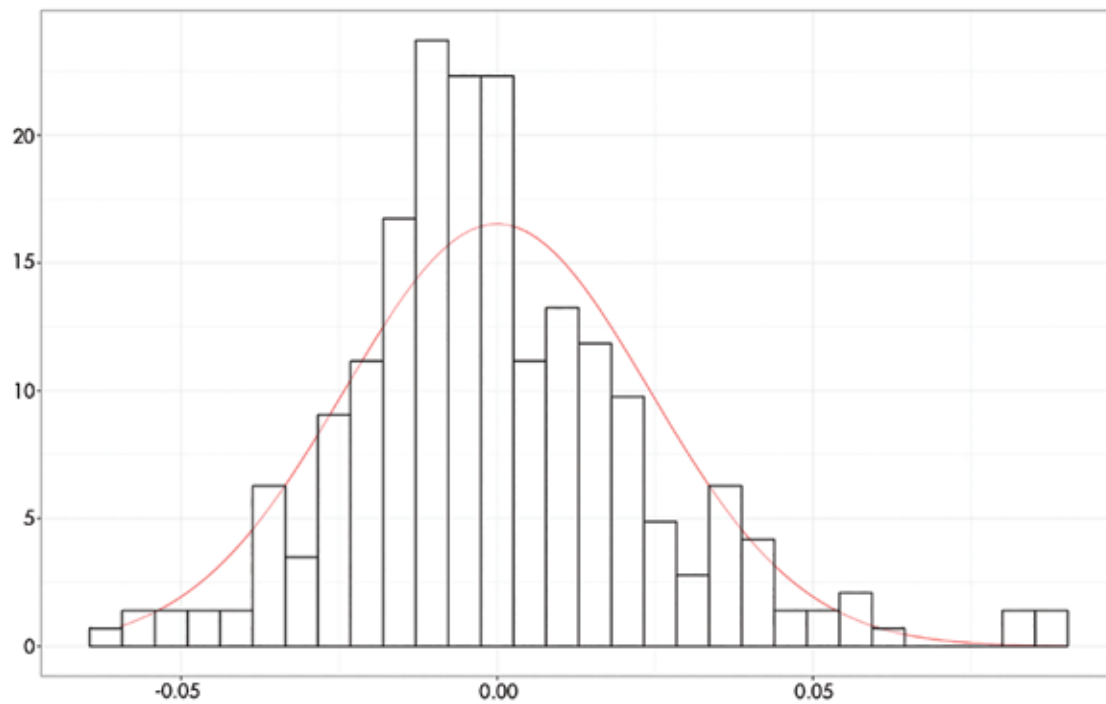


Figure 5. Random player coefficients (histogram).

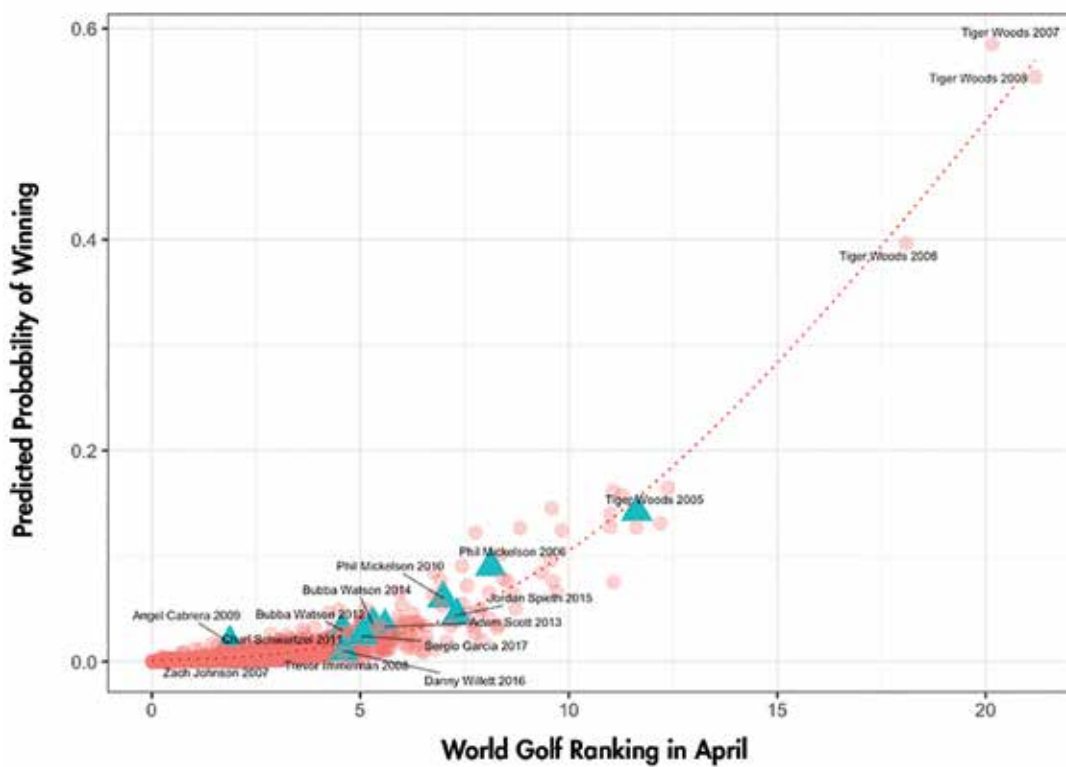


Figure 6. Probability of winning vs. OWGR.

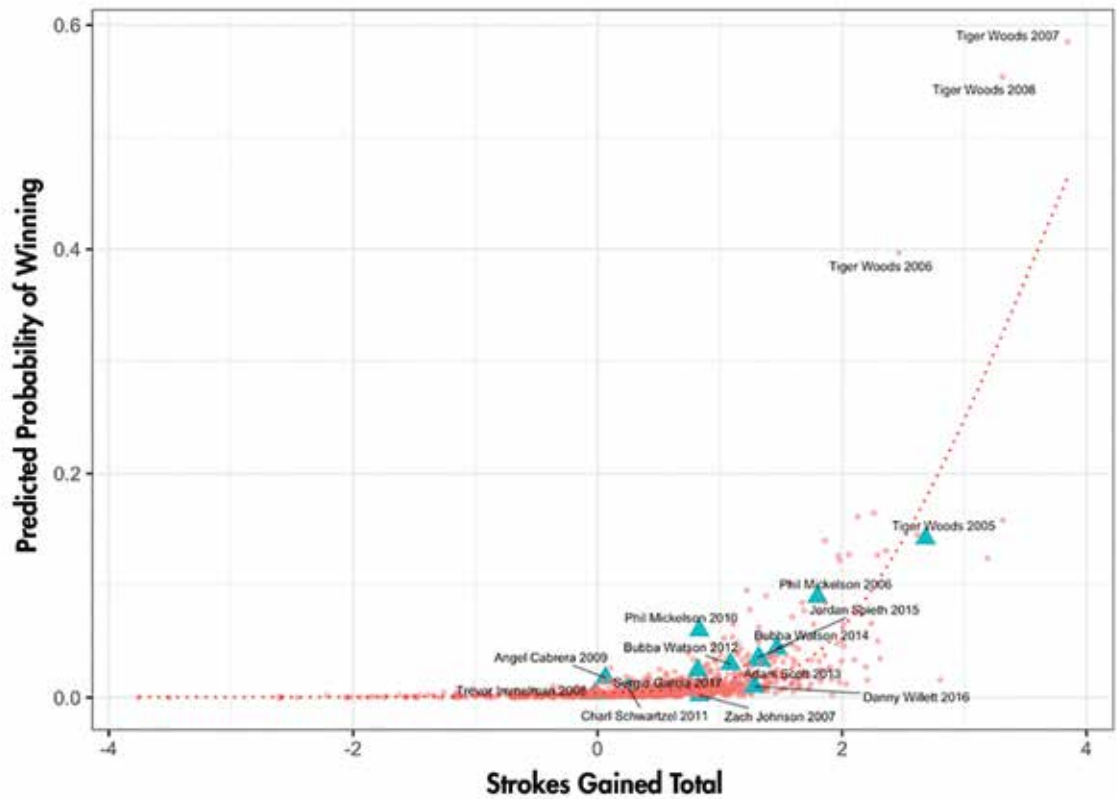


Figure 7. Probability of winning vs. Strokes Gained total.

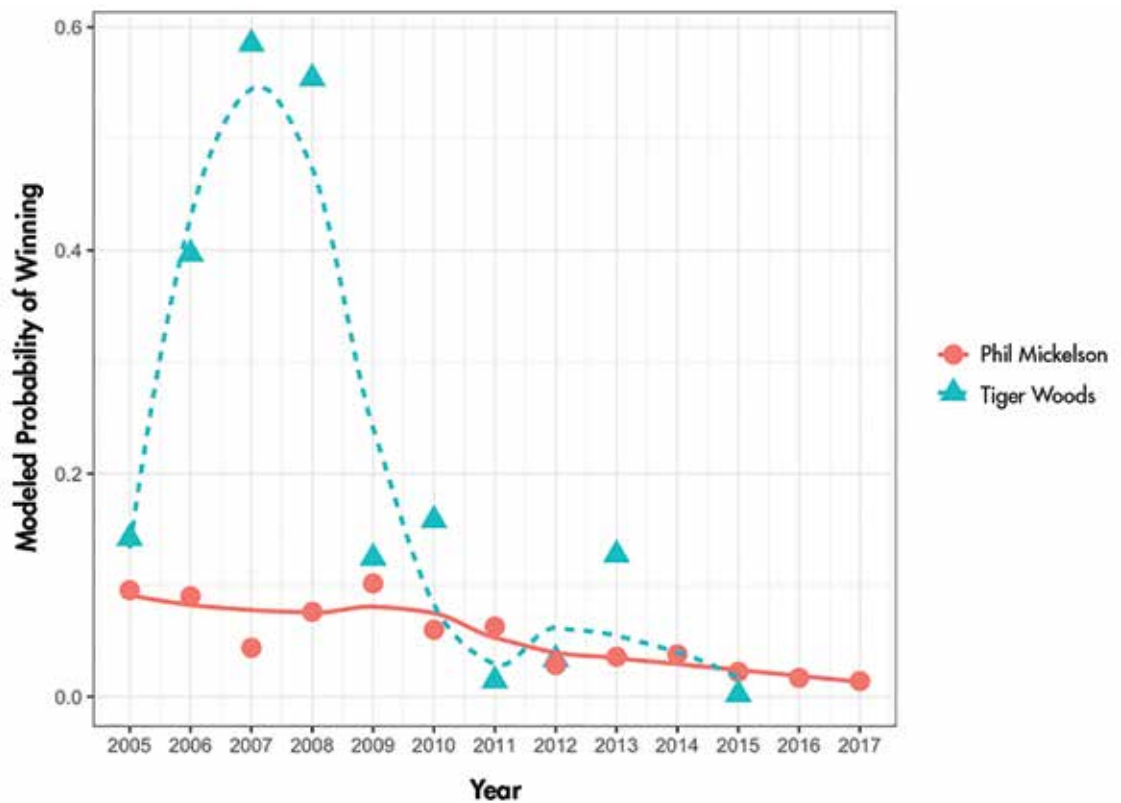


Figure 8. Probability of winning the Masters: Woods vs. Mickelson.

**Table 7—2016 Spieth Simulation Results**

Player	Score	Tee Box	SG Total	OWGR Pts	P(winning)
<b>Jordan Spieth</b>	<b>-7 or 245</b>	<b>10</b>	<b>1.86</b>	<b>11.000</b>	<b>0.9458</b>
Smylie Kauffman	+1 or 253	10	0.876	2.395	0.0004
Hideki Matsuyama	+2 or 254	10	1.378	4.518	0.0004
Dustin Johnson	-1 or 255	11	1.44	6.379	0.0135
Jason Day	E or 256	11	2.259	12.374	0.0062
<b>Danny Willett</b>	<b>-2 or 258</b>	<b>12</b>	<b>1.28</b>	<b>4.595</b>	<b>0.0247</b>
Lee Westwood	E or 260	12	0.052	1.977	0.0023
Soren Kjeldsen	-1 or 262	13	0.124	2.446	0.0067

Note: Matsuyama was playing in the group ahead of Spieth and had not completed the 10th hole. Since simulations cannot start a player in the middle of a hole, he was started on the 10th tee.

in blue. The curves were fit in R using the smoothing method “loess,” and show the overall trend of the data. World ranking seems to be the better predictor of Masters performance, due to the tighter and more-discernable pattern. Both figures show that probability of winning begins to increase dramatically for the very top players.

Figures 6 and 7 also show that Tiger Woods’s probabilities of winning during 2006 (0.397), 2007 (0.585), and 2008 (0.554) were truly remarkable. His strokes gained values and world rankings were so high during these years that he actually had a better chance of winning than the rest of the field. In 2007 and 2008, his strokes gained total were 1.5 shots better than the second best player on the PGA Tour per round. His world ranking points were over 20, which was more than twice as much as the second-ranked player.

When TV commentators used to ask each other if they would take Woods or the field, it was a reasonable question. Despite Woods’s extremely high probability of winning the Masters in that three-year stretch, he was not able to do it. He actually finished tied third in 2006, tied second in 2007, and second alone in 2008. For comparison, Figure 8 charts Woods’s and Phil Mickelson’s probabilities of winning through the years. Note a steady but slight decline for Mickelson as he ages into his 40s, and the dominant years from 2005 to 2010 for Woods.

Finally, we get back to Spieth in 2016. When Spieth walked to the 10th tee in the final round of 2016, what was his probability of winning? The simulation was run in “real time” with each player starting on the hole they were playing when Spieth moved to number 10. Table 7 lists the players’ scores and

starting positions, strokes gained, and world ranking, and probability of winning based on 100,000 simulations.

According to the model, Spieth had a 95% chance to win. A playoff would have resulted only 4% of the time. As golf fans know, Spieth proceeded to bogey 10 and 11, and then hit the ball in the water twice on the par three 12th and take a quadruple bogey seven. This opened the door for Willett, who did not let his chance get away.

The authors certainly do not know what was going through Willett’s mind when he suddenly found himself in the lead after Spieth’s meltdown on 12, but the pressure of holding the lead himself did not faze him. He played the last six holes in three under par, and won comfortably by three shots over Spieth and Westwood. He went from a two or three in a 100 chance with seven holes left to the green jacket.

A similar simulation was performed for the 2017 Masters. Sergio Garcia finally broke through and won his first major in a playoff with Justin Rose. However, on the 13th hole in the final round, Garcia trailed Rose by two shots. Rose hit a perfect tee shot, setting himself up for an eagle or birdie opportunity at the reachable par five. Garcia meanwhile hit his tee shot left of Ray’s creek into a bush and had to take a penalty drop.

It would have been fun to start a simulation from this point, but a limitation of this model is that it has to start players at the beginning of a hole. Garcia was able to pitch out of the trees down the fairway, and then pitch onto the green and make his putt for a par.

Meanwhile, Rose hit his second shot to the back of the green, but it took him three shots to get down,



**Table 8—2017 Simulation: Garcia vs. Rose**

Player	Score	Tee Box	SG - Total	OWGR Pts	P(winning)
Justin Rose	-8 or 260	14	1.148	4.490	0.83245
Sergio Garcia	-6 or 262	14	0.817	5.043	0.16755

so he also made a par. Despite Garcia managing to avoid losing any more ground, he was still two shots behind with five holes to play. What were his chances of winning from this point? The simulation results from 100,000 repetitions indicate that he still only had a 17% chance to win. A playoff between the two would have happened 11% of the time.

Looking again at Garcia and Rose from their tee shots on 13, a reasonable result would be a bogey for Garcia and a birdie for Rose. If Garcia went to the 14th tee four shots behind, the probability of him winning was only 0.033 in 100,000 simulations. This just emphasizes how important Garcia's great par save was to keeping him in the tournament.

## Conclusion

This study has shown that ordinal logistic regression can be used to effectively estimate probabilities of different scores on each hole for each player in the Masters golf tournament. Cross-validation analysis shows that the model has similar predictive abilities to the pre-tournament odds listed in Las Vegas. Predicting scores by hole allows for interesting real time, in-round simulations that cannot be conducted with a round-by-round predictive model.

Amongst the covariates in the model, the official world golf rankings appear to be the best single predictor of Masters performance. The model also

showed that using strokes gained statistics, long-game skills on the PGA Tour of driving and approach shots are better predictors of Masters success than the short game skills of chipping and putting. In addition, including a random player effect in the model allows for estimating which players tend to play the Masters better or worse than their skill level would indicate.

Future researchers interested solely in pre-tournament predictions might want to experiment with a round-by-round model and see if they get better results. Other options might be to make some adjustments to the covariates. For example, instead of using the prior 12 months of strokes gained data, the prior 24 months could be used. This might be helpful in getting more-accurate predictions for the European players who do not play as much on the PGA Tour.

A weighting system could also be used, with strokes gained statistics in more recent tournaments weighted more heavily.

The best improvement could be made by including pin positions and wind conditions. While this information was not available for this study, a historical record of pin locations and wind conditions probably does exist, or at least could easily be recorded in future years. ■

## Further Reading

- Agresti, A. 2002. *Categorical Data Analysis*. Wiley Series in Probability and Statistics.
- Arkes, J. 2016. The Hot Hand vs. Cold Hand on the PGA Tour. *International Journal of Sports Finance* 11, 99–113.
- Baldson, E.M. 2013. Risk Management with Tournament Incentives. *Journal of Quantitative Analysis in Sports* 9, 301–317.
- Clark, R.D. 2005. Examination of Hole-to-Hole Streakiness on the PGA Tour. *Perceptual and Motor Skills* 100, 806–814.
- McHale, I.G. 2010. Assessing the Fairness of the Golf Handicapping System in the UK. *Journal of Sports Sciences* 28(10), 1033–1041.

## About the Authors

**Erik Heiny** is an associate professor of mathematics at Utah Valley University. He received a bachelor's degree in both mathematics and statistics in 1992 from Colorado State University, a master's degree in statistics from Michigan State University in 1994, and a PhD in applied statistics from the University of Northern Colorado in 2006.

**Cody Frisby** is currently a data analysis engineer at IM Flash in Lehi, Utah. He received a bachelor's degree in statistics from Utah Valley University in December 2017.

# Defying the Odds: How Likely Are We to See Another Team Pull a 'Leicester' and Win the EPL?

*Craig A. Heard and A. John Bailer*

In 2016, sports fans witnessed the ultimate Cinderella story when Leicester City achieved the unimaginable: emerging as champions of the English Premier League (EPL) despite their 5,000:1 odds offered at the beginning of the season.

Before the start of the 2015–16 season, only five clubs had been crowned champions in the league's 24-year history. This unprecedented achievement by Leicester City motivated our investigation into how the teams of the EPL transition from season to season; whether a club's performance in one season is predictive of performance in the following season; and ultimately, how 5,000:1 odds could be justified using this information. Although many may see Leicester City's success as a fluke, this result also raises questions about which teams regularly make the top tier of the EPL, and which teams are left battling for survival.

Performance in the EPL has been evaluated using different game measures, monetary factors, and past performance. Beyond the obvious of scoring more goals than opponents and winning more games and a correlation between expenditure and record, nothing provides a formula for what it takes to be successful in the EPL. This is mainly due to the unique

continuity of the game, which makes it much more difficult to analyze when compared with other sports, such as baseball and American football.

One proposal for why this sport is so difficult to analyze comes down to the conclusion that the outcome of a game is split between skill of the team and luck. As former players and coaches, and current fans, we can attest to saying on more than one occasion, "We didn't deserve to win that game" or "We got lucky with that one." This is why analyzing football data should be approached with caution, but also why Leicester City's achievement stands out so much: They didn't just win a "one-off" game.

To investigate Leicester City's achievement, performance patterns of teams based on initial rankings (position in league rankings at the start of the season) must be considered. We conducted a Monte Carlo simulation study to see whether a simple simulation model could reflect an observed impact of initial rankings on final rankings (ranked final position at the end of the season). Using these results, we then estimate the probability of Leicester City (#14 ranked at the finish of the previous EPL season) or any other preseason ranked team winning the league.

## DEFINITIONS AND EXPLANATIONS

**EPL:** English Premier League

**Champions League:** most prestigious competition for teams ranked in the top 4 at the end of the season.

**Europa League:** teams ranked 5–7 at the end of the season

**Relegation:** teams ranked 18–20 at the end of the season.

**Goal Difference:** number of goals scored by one team minus the number of goals scored by its opponent

**Rankings at start of season for newly promoted teams:** 18 for Championship winner, 19 for Championship second place, and 20 for Championship playoff winner.

## Initial and Final Rankings for Premier League Clubs from 1996–2016

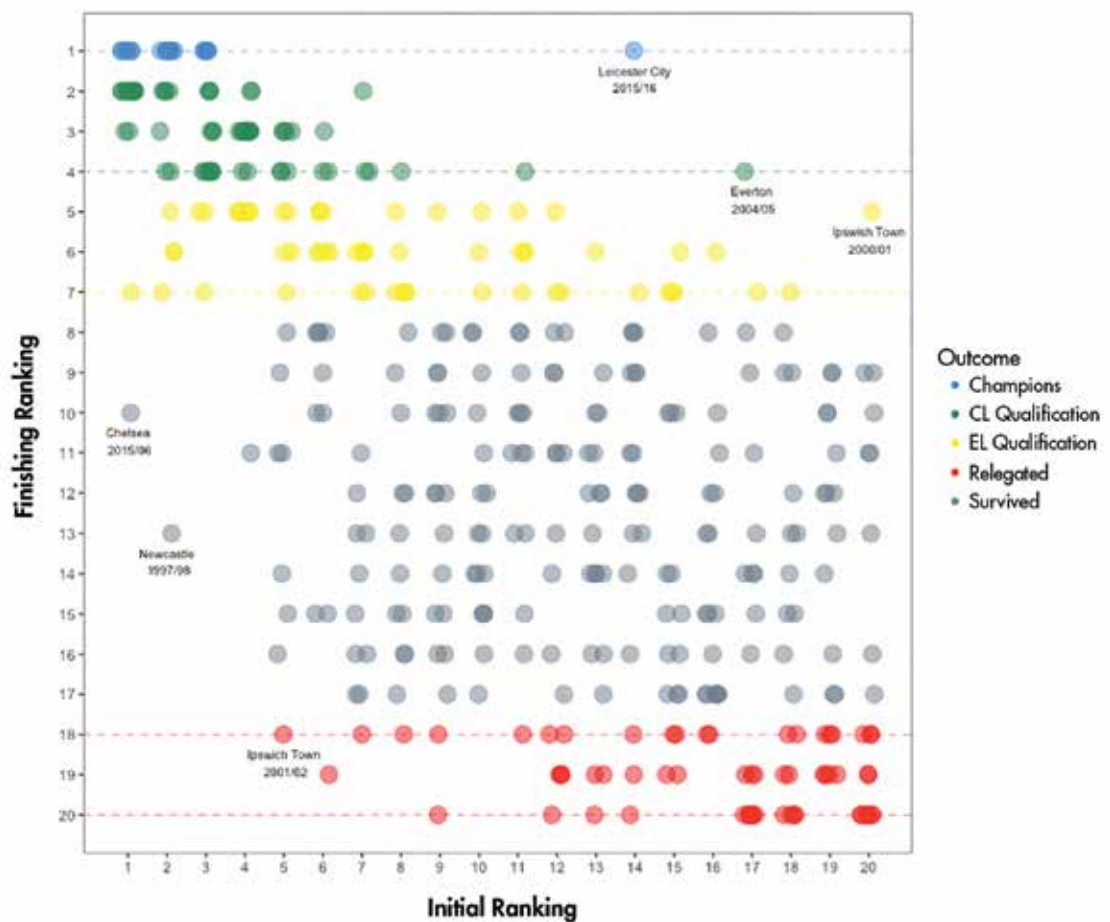


Figure 1. Initial ranking (jittered) against final rankings plotted for EPL clubs from 1996–2016.

### Are Final Season Rankings Related to Initial Rankings?

Initial rankings of the teams compared to their final rankings at the end of the season are illustrated in Figure 1; the plot highlights how, until the 2015–16 season, we had never seen a team with an initial ranking outside of the top three win the EPL in the last 20 years. We can also see that there have only been two teams outside of the top eight that have secured Champion League qualification

(finished the season in one of the top four positions).

The largest difference in the EPL's history is Ipswich Town, who were promoted through the playoffs from the Championship (league below the EPL) in the 1999–2000 season and then finished fifth in their first season in the EPL, a rise of 15 places. Unfortunately for their supporters, they also tie the record for the biggest drop by any team, losing 13 places and getting relegated the following season. Leicester City dropped

11 spots and finished 12th in the EPL in the 2016–17 season.

Teams are relegated from the EPL to the championship if they finish in the bottom three positions at the end of the season. These teams are then replaced by the teams finishing in the top two positions in the championship and the team that wins the championship playoff for the start of the following season.

Figure 2 explores the distribution of points for initially ranked teams. In the EPL, a team receives 3 points for a win, 1 point for

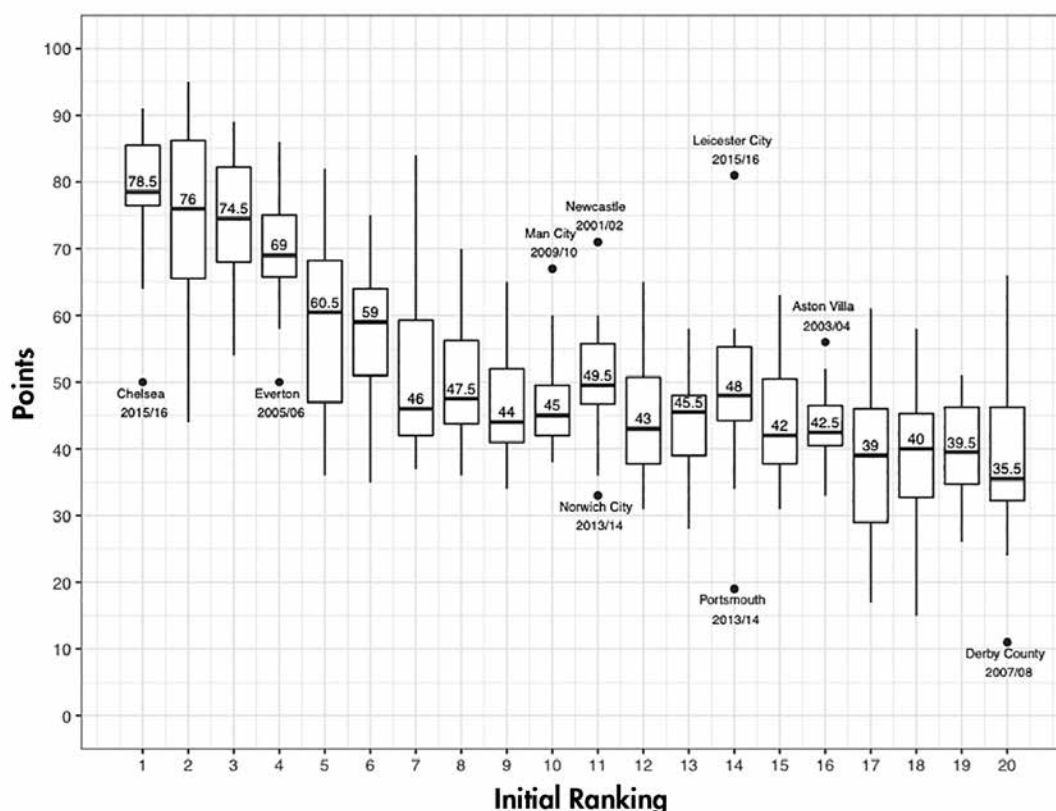


Figure 2. These box and whisker plots show the distribution of points for initially ranked teams for EPL clubs from 1996–2016. Outliers are highlighted with the club that had those points and the season they achieved those points.

a draw, and 0 points for losing the game.

What stands out the most here is how the top three initially ranked teams have a similar median indicating that these three initial positions seem to separate the elite clubs from the rest. Higher points total translate directly into increasing the chances of winning the league. Even more interesting is that there is a relatively small difference among the distribution of points between teams ranked 7 to 20.

This difference reveals the idea that no one is really safe outside the top six when it comes to surviving an EPL season. It has been commonly regarded by EPL managers who are battling it out for survival

that 40 points are enough to survive in the EPL. There have only been three occasions where a team has been relegated with 40 or more points. Figure 2 also highlights the teams that had unusual seasons; again, Leicester City appears to be a common outlier in both figures.

Complementary to Figure 2, goal differentials (goals scored minus goals conceded) during the last 20 years of the EPL were investigated by looking at each initial ranking (plot not shown). This pattern of the top three clubs separating themselves from the rest of the league was observed here as well. Clubs that have a beginning ranking in the top half of the table (#1–#10 at the start of the season) have a goal difference greater than

-10 ( $> 25$  for teams #1–4), while teams in the bottom half (#11–20) have a goal differential lower than -10. Teams with an initial ranking of 4 or higher are more likely to win by two or three goals when playing lower-ranked opposition, while clubs that are fighting for survival might only be clinching games by one goal.

That is not to say that lower-ranked teams do not beat higher-ranked teams because the data show that on any given day, it is possible for even a low-ranked team to defeat a much more-prestigious opponent. Nonetheless, to see a lower-ranked team completely embarrass a higher-ranked team by three or four goals is highly unlikely, according to the

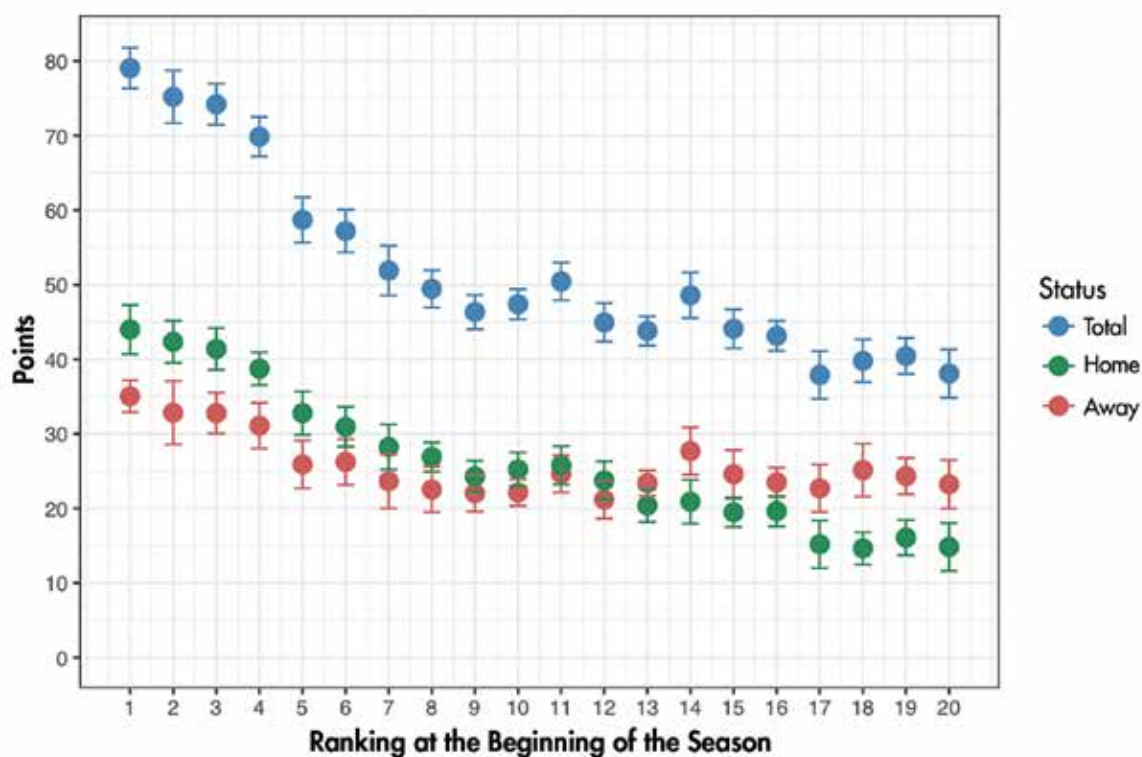


Figure 3. This plot shows the average points (home and away) gained by each initially ranked team for EPL clubs from 1996–2016. The error bars show two standard errors away from the points average for an initially ranked club.

data (with the exception of Leicester City in the 2015–16 season).

Figure 3 also adds to the evidence of this clear separation between the top four initially ranked teams and the rest of the league. Teams with a better away record than home record generate more points in away matches versus home matches if initially ranked lower than 12. Most sports have “home field” advantages, but the average points from the data collected do not support this theory, with respect to the EPL play of the lowest-ranked teams. The gap between the top four initially ranked teams’ home and away average points is quite large, and even the fifth and sixth initially ranked teams have a clear advantage at home.

The big surprise is that the bottom four initially ranked teams

all have poorer records/generate fewer points at home than away from home. This could be due to clubs in the top half of the table coming to these teams and historically gaining at least one point from the game (either winning or drawing the game). It is also clear from the chart how much of an outlier Leicester City’s season was in 2015–16 with its average points being much higher than could reasonably be predicted. Leicester City had 81 points in the 2015–16 season when #14 pre-season ranked teams were expected to earn between 18 and 24 points at home and between 25 and 32 away, totaling between 43 and 56 for the season.

The probabilities of winning are plotted versus the difference of strength between initially ranked teams in Figure 4. This plot is

faceted for team initial rankings strata 1–4, 5–6, and 7–20. These probabilities were derived empirically—obtained from the previous 19 seasons (19 seasons before Leicester’s title-winning year) where a higher-ranked team  $i$  played a lower-ranked team  $j$ . The difference is simply team ranked  $i$  minus team ranked  $j$ . A difference of 13 in strata Ranks 1–4 would represent the probability of winning for teams initially ranked 1 vs. 14, 2 vs. 15, 3 vs. 16, and 4 vs. 17.

A home field advantage can clearly be seen when it comes to the probability of winning for higher-ranked teams against lower-ranked teams. The probability of a higher-ranked team winning increases (at both home and away locations) as the difference between the team’s initial rankings increases for teams initially ranked 1–4. This



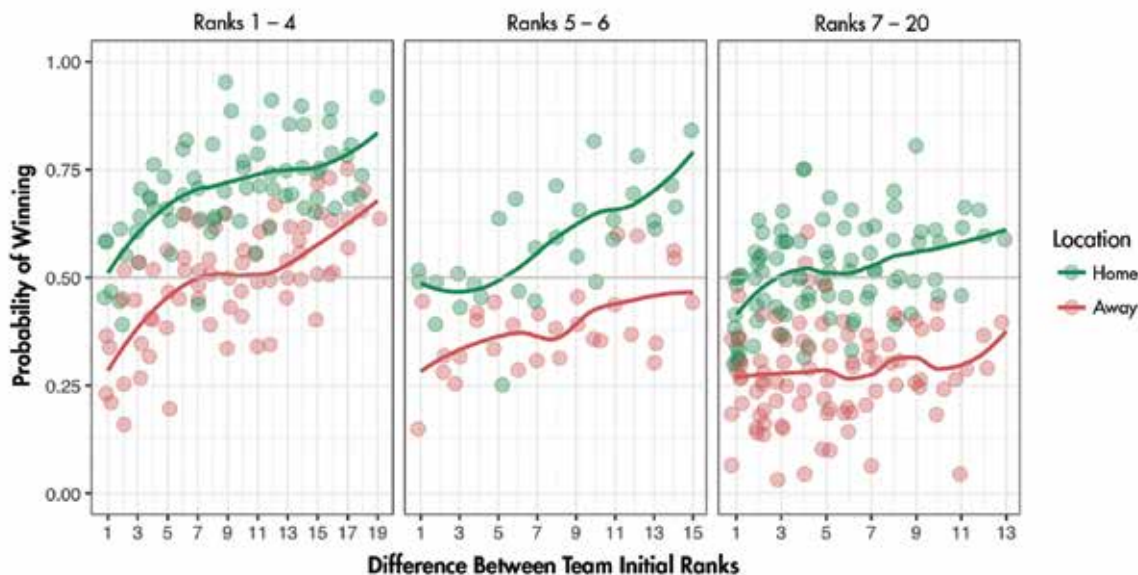


Figure 4. This plot shows the probability of winning for a team playing against a lower-ranked team based on the difference between their initial rankings.

is also the case for teams initially ranked 5–6, but the difference in rankings for teams initially ranked 7–20 exhibits a slight increase as the difference in rankings among teams increases.

This again provides support for how close in quality teams are with initial rankings 7 to 20. The difference in team strength does not have an effect on the probability of drawing when they play teams that are in the same stratum and lowest-ranked teams.

## Simulating EPL Seasons

An EPL season comprises 380 games, with each of the 20 teams playing every other team twice in a season, at home and away. Each of these games results in one of three outcomes—Win, Draw, or Loss—with a point value of 3, 1, and 0, respectively. The team with the most points at the end of the season is crowned EPL Champion. A trinomial probability distribution is the foundation

for simulating a game. A season is generated from evaluating 380 such trinomial experiments.

For the purpose of this study, estimated probabilities for the outcomes of every game during a season had to be created (380 total games) based on the 19 seasons (1996–2015), i.e., excluding Leicester City’s miraculous season. Probabilities were derived empirically, obtained from the previous 19 seasons of competition where the team ranked  $i$  played the team ranked  $j$  at location  $k$ . Therefore, home advantage was considered in the set of trinomial probabilities. At least two of the trinomial probabilities were greater than 0. Data were also collected on the initial ranking, final rankings, and final goal differentials for the same time period.

In the EPL, it is possible for teams to be tied with the same number of points at the end of the season. To break this tie when determining final rankings, each team for each season was assigned a goal differential once their points

had been calculated. This goal differential was sampled from the pool of 19 goal differentials collected for each initially ranked team, and the goal differential was used to separate teams that finished with the same number of points.

Once the goal differential was assigned, the teams were then ranked in final position. For example, if a team started a season with an initial ranking of 5 then their goal differential at the end of the simulated season would be randomly selected from the pool of 19 goal differentials for a team initially ranked as 5.

Finally, we generated a set of 10,000 seasons. Once all the seasons were simulated, we created tables for each season and assigned rankings for final positions to each team.

Generating simulated seasons based on our data from multiple EPL seasons allows us to estimate probability or odds of Leicester City winning the league in 2015–16, given their 14th-place finish in the previous season.

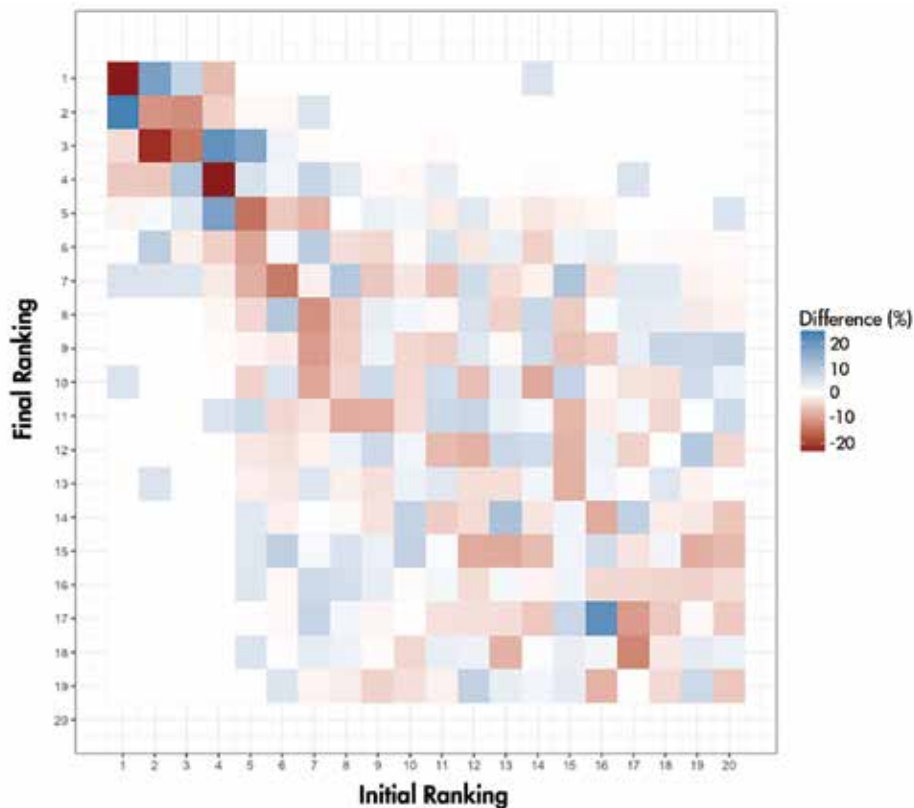


Figure 5. Heat map showing the difference between the empirical and simulation proportions of where initially ranked teams finished.

### Can Simulated EPL Seasons Reflect the Observed EPL Seasons?

Figure 5 shows a heat map displaying the difference between the empirical probabilities and simulation probabilities of where initially ranked teams will finish. Differences can be seen between the simulation results and the empirical results, when comparing the initially ranked teams 1–4. The simulation results and empirical results are very similar for initially ranked teams 5–20, indicating how similar these clubs actually are when it comes to performance in the league.

It is clear that the biggest difference appears among the top four initially ranked teams. Looking at the empirical probabilities (not shown in this paper) among these

top ranked teams, no game has a probability of winning greater than or equal to 50%. This demonstrates that a victory in one of these games is crucial in determining who wins the EPL. Winning the majority of these games is crucial to winning the league, since the separation of points between the top three teams is so close.

The simulated proportion looks similar to that of the empirical proportions from the collected data; however, there are vast differences when it comes to the teams ranked fourth or higher. In particular, the simulation proposes that a team with an initial ranking of 1 will retain the title 54% of the time, which is 24% higher than the empirical probabilities we have. As a consequence of so many first-place finishes, a team initially ranked first also had 18% fewer second-place finishes

for predicted versus observed because of the higher number of retained titles.

Additionally, a team initially ranked in second place finished a season in first place less often in the simulation compared to the empirical probabilities (24% vs. 50%).

The overestimation for the proportion of seasons that the first initially ranked team retained its title in the simulation raises a few questions. Between EPL seasons, a lot of events can occur such as transfers in, transfers out, new staff, and occasionally a new owner pumping millions of capital into the club. Perhaps finishing #2 or #3 results in a greater investment or motivation to achieve the winning the EPL in the next season.

There also seems to be something unique about winning the previous season when entering the

**Table 1—Simulated Probabilities and Estimated Odds of Where Initially Ranked Teams Will Finish Calculated Based on 20 Sets of Simulations (10,000 seasons per simulation)**

Initial Rank	Final Rank	Simulated Probability	Standard Error	Odds Against
#14 (LC)	#1	0.000015	0.0000082	66,666:1
#7–20	#1	0.000385	0.0000078	2,596:1
#1 (Chelsea)	#10	0.000025	0.0000099	39,999:1
#1–3	#10	0.000650	0.0000064	1,526:1

next season. Every other team is out to beat you. Clubs facing the previous season's champion may have extra motivation to succeed and possibly create an upset in the game.

When Chelsea won the league in 2014–15 by 8 points, one fan make the statement on BBC 5 Live that “Chelsea have outgrown the league.” This couldn't have been more wrong, because Chelsea slumped to the lowest finish by a team initially ranked first in the following season.

After winning a title, it is safe to say that retaining the title is more difficult than the simulation indicates. Over the last 20 years, most EPL seasons see at least one other challenger for the title and sometimes two or three. At the beginning of the season, several teams are aiming to claim the title, which explains why we have seen teams initially ranked second and third winning the league eight and five times, respectively, out of 20 seasons.

Interestingly, no team that was initially ranked fourth place or higher ever got relegated, although there were incidences of fifth place and sixth place getting relegated.

In one set of 10,000 simulated seasons, one team initially ranked 14th at the beginning of the season won the league (10,000:1 odds against). In total, four teams outside the top six won the league

(7, 9, 11, and 14), with initially ranked team 11 winning three out of the 10,000 seasons! This yields odds against any team ranked 7–20 winning the league of .9993/.0007 of ~1,427:1.

By looking at the earlier discussion of the distribution of points and goal differentials, it appears that the clubs that fall in this range have just as much chance of finishing in the top half of the league as finishing in the bottom half, are just as well likely to be relegated, and might even be able to go the whole way in “Doing a Leicester.” This leads us to believe that the odds only given to Leicester at the beginning of year were, in fact, quite conservative.

The results in Table 1 show that the simulation estimates odds of 66,666:1 for a team initially ranked 14th in the league, which is pretty ridiculous. However, our analysis provides strong support for teams outside the top six being in a similar stratum, so by combining these teams, we can see the odds improve significantly and more closely match the odds given by the bookies at the beginning of the 2015–16 season. We also can see odds just as large for a team initially ranked 1 dropping down nine places to 10th to just under 40,000:1. This reduces if we look only at the stratum of teams #1–3 drop to just over 1,500:1.

## How Should you Place your Bets on Future EPL Season Outcomes?

Within the last 20 years, we have seen the amount of money being invested in the league and clubs skyrocket. Promotion into this league brings in huge amounts of revenue for television rights, among other endorsements. The EPL is where wealthy businesspeople invest in a club and spend billions of pounds to improve facilities, increase stadium capacity, and—more importantly—provide the capital needed to make transfers. Although football is a team sport, there have been cases where signing one or two players during the January or summer transfer window has had a huge effect, either during the season or in preparation for the following season.

There are natural extensions to this project. First, this method could be applied to other football leagues around the world, such as La Liga and the Bundesliga. This model also could be applied to different sports. How would investigation into initial rankings fare in the NFL, NBA, or Olympics? Could this work in a World Cup format? Could end-of-the-season form from the previous season be incorporated into a study of performance in the next season?

Two teams with a similar final ranking may be very different in terms of next-season potential based on how they finish the current season. For example, finishing a season on a winning streak versus losing most of the final games is likely to suggest two teams with different potentials at the start of the next season.

The form of both players and the team itself is another implication of our model, since teams will go in and out of form throughout the season. The main focus of this research was looking at Leicester City's season. We know that in the 2014–15 season, they won seven of their last nine games when they were almost certainly facing relegation. This momentum continued into the following season. Further exploration into runs testing might be a tool to use in the future, along with giving more-recent seasons more weight than more-distant seasons.

## Did We See a Miracle Season in Sport?

At the beginning of the 2015–16 EPL season, Leicester City was given 5,000:1 odds of winning the EPL. Through simulation, we found this to be a fairly reasonable estimate when considering that teams outside the top six teams are roughly equal when it comes to their performance in the league, meaning any of

these teams have similar chances to pull a “Leicester.”

Our simulation results calculate the odds for a team like Leicester City's initial ranking of 2,596:1. However, due to limitations of the simulation, such as independence violations, as well as the other lurking variables that make football so difficult to analyze, this estimation of Leicester City winning the league at the beginning of the season could be argued to be not a bad guess, especially given how difficult it is to estimate very small probabilities.

In addition to the dramatic rise of Leicester City, the results highlight that Chelsea's fall from Champion to #10 at the end of the 2015–16 season was equally unusual, but didn't garner the headlines.

Although the EPL has a relatively short history, it has been noticeable that three to four teams have dominated the top of the league table over the last 20 years. Visual explorations of these seasons support this separation between the “top” clubs and the rest of the league not only in terms of point distribution, but also in the distribution of goal differentials. We have discovered not that this gap exists, but the size of the gap and the closeness between the other clubs in the league. The results provide a great insight into the stratification of this league and creates hypotheses that should be explored for future studies.

In addition to the data visualizations, we created a Shiny App using R Statistical Software that allows users to interactively explore the history of any of the clubs that have competed in the EPL since 1996 over the next 20 years. This app demonstrates the volatility in performance for a lot of the teams, but also shows how consistent some teams have been throughout the last 20 years (<http://dataviz.miamiob.edu/EPLClubExplorer>).

Here, a simple computer simulation of the EPL was implemented rapidly to provide an initial exploration of initial rankings on season performance for any sport with league structure, although we suspect that generalization to other football leagues would be the easiest application to consider as a next step. Further, making season performance data easily explored and visualized will lead to greater engagement by fans.

Based on this investigation, the odds that *some* lower-ranked team winning the league was higher than 5,000:1, but for a *specific* team ranked outside the top 6, 5,000:1 odds were better for the bookies. However, this makes sense for the bookies because if every team outside the top six had odds of winning the league in the range of 50,000:1, a lot more people would have made small bets. We believe that Leicester City's result was actually more impressive than the bookies' initial 5,000:1 odds suggested and that yes, they did in fact defy the odds.

To come back to the question of *How likely are we to see another team pull a 'Leicester' and win the EPL?*, our answer is “not very.” We are excited to have been alive to see the “Leicester miracle”; however, neither of us is placing our bets on happening again soon (unless the bookies increase the odds against dramatically). ■

## Further Reading

- Din-Houn Lau, F., and Gandy, A. 2016. Enhancing football league tables. *Significance*, 8–9.
- Gandy, R. 2016. Second season syndrome. *Significance*, 26–29.
- Oberstone, J. 2009. Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success, *Journal of Quantitative Analysis in Sports* 5:3.

## About the Authors

**Craig Heard** is a market analyst at JP Morgan Chase. He earned an MS in statistics from Miami University in 2017. More significantly, he is a lifelong football player, ex-collegiate athlete, and dedicated Manchester United fan.

**John Bailer** is a University Distinguished Professor and chair of statistics at Miami University. After growing up playing soccer, he coached youth soccer for many years. He also serves as a regular panelist on the Stats+Stories podcast ([www.statsandstories.net](http://www.statsandstories.net)). He earned a PhD in biostatistics from the University of North Carolina in Chapel Hill in 1986.



## Code of Silence

*How private companies hide flaws in the software that governments use to decide who goes to prison and who gets out*

Rebecca Wexler

One day in early January, a letter appeared on my desk marked DIN92A5501—an inmate’s identification number from the Eastern Correctional Facility in upstate New York. The author, Glenn Rodríguez, had drafted it in upright, even letters, perfectly aligned. Here, in broad strokes, is the story he told:

Rodríguez was just 16 at the time of his arrest, and was convicted of second-degree murder for his role in an armed robbery of a car dealership that left an employee dead. Now, 26 years later, he was a model of rehabilitation. He had requested a transfer to Eastern, a maximum-security prison, to take college classes. He had spent four and a half years training service dogs for wounded veterans and 11 volunteering for a youth program. A job and a place to stay were waiting for him outside. And he had not had a single disciplinary infraction for the past decade.

Yet, last July, the parole board hit him with a denial. It might have turned out differently but, the board explained, a computer system called COMPAS had ranked

him “high risk.” Neither he nor the board had any idea how this risk score was calculated; Northpointe, the for-profit company that sells COMPAS, considers that information to be a trade secret. But Rodríguez may have been stuck in prison because of it.

Proprietary algorithms are flooding the criminal justice system. Machine learning systems deploy police officers to “hot spot” neighborhoods. Crime labs use probabilistic software programs to analyze forensic evidence. Judges rely on automated “risk assessment instruments” to decide who should make bail, or even what sentence to impose.

Supporters claim that these tools help correct bias in human decisionmaking and can reduce incarceration without risking public safety by identifying prisoners who are unlikely to commit future crimes if released. Critics argue that the tools disproportionately harm minorities and entrench existing inequalities in criminal justice data under a veneer of scientific objectivity.

Even as this debate plays out, the tools come with a problem that is slipping into the system unnoticed: *ownership*. With rare exceptions, the government does not develop its own criminal justice software; the private sector does. The developers of these new technologies often claim that the details about how they work are “proprietary” trade secrets and, as a result, cannot be disclosed in criminal cases. In other words, private companies increasingly purport to own the means by which the government decides what neighborhoods to police, whom to incarcerate, and for how long. And they refuse to reveal how these decisions are made—even to those whose life or liberty depends on them.

The issue has been percolating through criminal proceedings for years. I work for the Legal Aid Society of New York City, defending criminal cases that involve computer-derived evidence. I regularly see defendants denied information that they could use to cross-examine the evidence against them because of a trade secret.

*This article originally appeared in the June/July/August issue of the Washington Monthly (<https://washingtonmonthly.com/magazine/junejulyaugust-2017/code-of-silence/>) and is reprinted here by permission. Minor edits have been made to conform with house style.*



At presstime for this issue of *CHANCE*, in *Loomis v. Wisconsin*, the U.S. Supreme Court was deciding whether to review the use of COMPAS in sentencing proceedings.

Eric Loomis pleaded guilty to running away from a traffic cop and driving a car without the owner's permission. When COMPAS ranked him as high risk, he was sentenced to six years in prison. He tried to argue that using the system to sentence him violated his constitutional rights by demoting him for being male. But Northpointe refuses to reveal how it weights and calculates sex.

On June 26, 2017, the Supreme Court decided not to review the use of COMPAS in sentencing proceedings in the context of the *Loomis v. Wisconsin* case, and denied the petition for certiorari in that case. The court is likely to face more petitions for certiorari on related issues in similar cases.

We do know certain things about how COMPAS works. It relies in part on a standardized survey where some answers are self-reported and others are filled in by an evaluator. Those responses are fed into a computer system that produces a numerical score. But Northpointe considers the weight of each input, and the predictive model used to calculate the risk score, to be trade secrets.

That makes it hard to challenge a COMPAS result. Loomis *might* have been demoted because of his sex, and that demotion *might* have been unconstitutional, but as long as the details are secret, his challenge can't be heard.

What surprised me about the letter from Eastern was that its author could prove something had gone very wrong with his COMPAS assessment. The "offender rehabilitation coordinator" who ran the assessment had checked

"yes" on one of the survey questions when he should have checked "no." Ordinarily, without knowing the input weights and predictive model, it would be impossible to tell whether that error had affected the final score. The mistake could be a red herring, not worth the time to review and correct.

Glenn Rodríguez had managed to work around this problem and show not only the presence of the error, but also its significance. He had been in prison so long, he later explained to me, that he knew inmates with similar backgrounds who were willing to let him see their COMPAS results.

"This one guy, everything was the same except question 19," he said. "I thought, this one answer is changing everything for me." Then another inmate with a "yes" for that question was reassessed, and the single input switched to "no." His final score dropped on a 10-point scale from 8 to 1. This was no red herring.

So what is question 19? The New York State version of COMPAS uses two separate inputs to evaluate prison misconduct. One is the inmate's official disciplinary record. The other is question 19, which asks the evaluator, "Does this person appear to have notable disciplinary issues?"

Advocates of predictive models for criminal justice use often argue that computer systems can be more objective and transparent than human decisionmakers. But New York's use of COMPAS for parole decisions shows that the opposite is also possible. An inmate's disciplinary record can reflect past biases in the prison's procedures, as when guards single out certain inmates or racial groups for harsh treatment. Question 19 explicitly asks for an evaluator's opinion. The system can actually

end up compounding and obscuring subjectivity.

That's what happened to Glenn Rodríguez. "It took a lot of energy and effort to maintain a clean record for the duration I had," he told me. Looking at his fellow inmates' COMPAS reports, he realized that some guys who had engaged in violent behavior within the past two years, but whose evaluators had checked "no," had gotten low prison misconduct scores.

Rodríguez went before the parole board last July. "This panel has concluded that your release to supervision is not compatible with the welfare of society," the board explained. Of significant concern was his "high COMPAS risk score for prison misconduct."

Trade secrets are a form of intellectual property for commercial know-how that are both stronger and weaker than patents. When the government grants a patent, no one is allowed to use your invention, period—but only for 20 years. After that, it's open season. By contrast, a trade secret lasts as long as you can keep it, well, secret. If rivals obtain the secret through "misappropriation" (that is, lying, spying, or fraud), you can sue them, and they can even face criminal charges. But if they reverse-engineer your product or simply come up with the same idea on their own, or if you just do a bad job hiding it, then it's too bad for you. Software developers like trade secrecy because their technology is not always patentable, and because patents are expensive to acquire and enforce.

There is debate among legal scholars about why the law recognizes trade secrets. Some even argue that it shouldn't. But the most commonly accepted rationale is that granting protections for information that may not be patentable, like an abstract idea

or a mathematical formula, will create an incentive for new intellectual creations.

What's alarming about protecting trade secrets in criminal cases is that it allows private companies to withhold information not from competitors, but from individual defendants like Glenn Rodríguez. Generally, a defendant who wants to see evidence in someone else's possession has to show that it is likely to be relevant to his case. When the evidence is considered "privileged," the bar rises: He often has to convince the judge that the evidence could be *necessary* to his case—something that's hard to do when, by definition, it's evidence the defense hasn't yet seen.

Based on state appellate court opinions, the invocation of trade secrets to prevent criminal defendants from accessing evidence against them didn't start happening frequently until the 1990s, when companies began refusing to disclose details about DNA testing kits that were being adopted by forensic labs around the country. There was some early pushback by judges and experts, but eventually, most courts ruled that DNA test kit manufacturers were entitled to keep aspects of their methods secret—and that prosecutors could still use the results as evidence to convict. In the past five years, courts in at least 10 states have ruled this way for DNA analysis software programs.

Private companies increasingly purport to own the means by which the government decides what neighborhoods to police, whom to incarcerate, and for how long. And they refuse to reveal how these decisions are made—even to those whose life or liberty depends on them.

Like any technology, though, DNA testing can be flawed.

In 2016, Michael Robinson, a death penalty defendant in Pennsylvania, tried unsuccessfully to subpoena the source code for a probabilistic genotyping software program called TrueAllele. A 30-year-old "family guy" with no prior criminal history, Robinson had been charged with murdering two people. TrueAllele matched his DNA to a bandana found near the scene of the crime.

Probabilistic genotyping software results are not gold standard DNA evidence. The programs were developed to test tiny amounts and complicated mixtures of DNA, and their accuracy is disputed. Last September, President Obama's Council of Advisors on Science and Technology found that more testing is needed to establish the validity of programs like TrueAllele.

Robinson sought to evaluate the TrueAllele code and check whether it worked the way its developer claimed. The judge denied his request. One reason she gave stands out: TrueAllele's developer, Mark Perlin, had said that ordering the code disclosed to the defense could "cause irreparable harm to the company, as other companies would be able to copy the code and put him out of business." As a result, the judge decided that compelling disclosure would be unreasonable. "Dr. Perlin could decline to act as a Commonwealth expert," she wrote, "thereby seriously handicapping the Commonwealth's case."

Robinson was forced to defend himself without access to the code. Despite the DNA test results, in February, he was acquitted on all counts. While the outcome was ultimately a happy one for Robinson, he had to sit through the trial knowing that the jurors were weighing evidence that he was unable to fully scrutinize and con-

test. How many future defendants will be wrongfully convicted based on misleading "proprietary" DNA software that they couldn't see or challenge?

Recently, companies have begun invoking proprietary secrets in the context of police investigatory tools. Accessing information about how these technologies work can be critical to a defendant's case. He or she might want to argue that they violate privacy rights, or aren't reliable enough to justify an arrest. In our adversarial legal system, these claims by individual defendants are often the main way to hold police to account.

That's what happened with Stingrays. A Stingray is a military surveillance device that masquerades as a cellphone tower to suck up information from your phone. When the manufacturer, the Harris Corporation, applied for certification by the Federal Communications Commission, it requested that information about the technology be kept secret for both law enforcement purposes and to maintain its commercial "competitive interests." As a result, police departments around the country signed non-disclosure agreements promising to conceal details about how the technology works—and even its mere existence—from defendants, courts, legislatures, and the public.

One man undid the secrecy scheme. When Daniel Rigmaiden was arrested for wire fraud and identity theft in 2008, he insisted that police must have used a secret device to beam "rays into his living room" and gather data about his location. After years representing himself *pro se* from a prison cell, he proved that he was right. He noticed a handwritten note with the word "Stingray" buried in his own 14,000-page court file.

Internet searches for the term yielded a Harris Corporation brochure and a purchase order for the device from the police department that had arrested him. Some courts have since found that warrantless use of Stingray devices violates the Fourth Amendment—holdings that would have been impossible without Rigmaiden's efforts. That means police spent years getting away with potentially unconstitutional Stingray searches and hiding their tracks with non-disclosure agreements.

Of course, not all secrecy is driven by profit. Some investigative methods must be kept under wraps to be effective. If anyone could predict IRS audits or airport security screenings, fraudsters and terrorists could avoid getting caught. The trouble is that the flip side is also true: Excessive secrecy can let police evade accountability for illegal or unconstitutional methods. Proprietary technology makes this too easy: First, outsource policing techniques to private companies. Then, claim those techniques are trade secrets.

The use of predictive policing tools shows how this can occur. These computer systems use machine learning to forecast where crimes are likely to be committed. One leading vendor, PredPol, has refused for years to reveal certain details about how its system forecasts future crimes. Wanting to protect that information is understandable: The tool took six years to develop and now generates an estimated \$5–\$6 million in annual revenue.

In January 2016, PredPol finally responded to public pressure and published a general description of its algorithm. That allowed independent researchers from the Human Rights Data Analysis Group to re-implement and test it. They showed that

applying the algorithm to police records could exacerbate past racially biased policing practices: Even when crimes were spread evenly throughout a city, PredPol would home in on areas that were overrepresented in police databases, intensify policing in those same areas, then use the foreseeable spike in crime reports to justify its earlier predictions.

The problem comes from the data, not the algorithm, but PredPol's describing the algorithm publicly helped researchers to demonstrate the issue empirically.

That doesn't mean we should never use machine learning systems; researchers are developing methods to audit, simplify, and try to reduce bias in predictive models. But it means that if courts allow the systems to be cloaked in secrecy, we may not be able to find the flaws, much less be able to fix them. What kinds of transparency we need most is a matter of technical debate. Trade secrets should play no role in determining the answer.

Once Glenn Rodríguez had figured out that a single survey response had swung his "prison misconduct" score from low to high, he sent a written complaint to a supervisor at Eastern. The "yes" response to question 19, he argued, was at odds with his exemplary behavioral record and was likely to hurt him at his next parole hearing, in January. Without COMPAS, his case for parole was nearly perfect. Question 19 was standing between him and freedom.

Rodríguez got farther than most people in his position. Thanks to the network of fellow inmates who shared their COMPAS scores, he was able to convince the rehabilitation coordinator that his score was inaccurate. "The question surrounding your disciplinary should be changed," she wrote in a letter to

Rodríguez last September. "Since you will be going to the Board in less than a year, we need to make sure the original one isn't used."

For Rodríguez, the next step was to wait. And wait. Despite the written assurances, no new COMPAS was provided. He sent letters to attorneys. (One arrived on my desk.) New Year's came and went. He filed a formal complaint with the Inmate Grievance Resolution Committee. The score was never fixed.

There's no question that we could use some innovation in the criminal justice domain. New technologies could help us find and convict criminals, exonerate the innocent, reduce human bias, and incarcerate fewer people, but recognizing the benefits of innovation does not require permitting developers to withhold their secrets from individual defendants.

It is one thing to argue that forcing companies to disclose trade secrets in public would hurt business and derail technological progress. It's another to claim that making them share sensitive information with the accused and their defense team, in the controlled context of a criminal proceeding, would do the same.

The most common justification for withholding proprietary information from a defendant is that without that guarantee, innovative companies will be deterred from investing in new criminal justice technology or from selling existing products to the government. But it isn't always clear that this concern is legitimate.

Take TrueAllele, the probabilistic DNA testing software. Mark Perlin, who developed it, did not answer my requests for an interview, but he has submitted declarations to courts across the country explaining that allowing defendants, their attorneys,

and defense expert witnesses to see his “source code” would enable the reverse engineering of the TrueAllele technology, allowing others to learn the trade secrets that keep [his company] solvent.” Prosecutors have fallen in line with Perlin’s view. One warned a court that ordering the code disclosed to a defense team would be “financially devastating.”

Perlin’s more-transparent competitors appear to be doing just fine, though. TrueAllele’s main rival, a program called STRmix, which claims a 54 percent U.S. market share, has an official policy of providing defendants access to its source code, subject to a protective order. Its developer, John Buckleton, said that the key to his business success is not the code, but rather the training and support services the company provides for customers.

“I’m committed to meaningful defense access,” he told me. He acknowledged the risk of leaks. “But we’re not going to reverse that policy because of it,” he said. “We’re just going to live with the consequences.”

Remember PredPol, the secretive developer of predictive policing software? HunchLab, one of PredPol’s key competitors, uses only open-source algorithms and code, reveals all of its input variables, and has shared models and training data with independent researchers. Jeremy Heffner, a HunchLab product manager and data scientist, explained why this makes business sense: Only a tiny amount of the company’s time goes into its predictive model. The real value, he said, lies in gathering data and creating a secure, user-friendly interface.

HunchLab is not alone... another start-up in the field, CivicScape, published its source code

and examples of its input variables online. Publishing models and the actual data used to train them would be even better, but the disclosure was a step in the right direction.

The fact that competitors in the same field as products like TrueAllele and PredPol have no problem revealing details about their methods should make courts less willing to take a company’s word when it comes to the need for total secrecy. It’s too easy for claims about financial devastation to mask a more troubling motive: avoiding scrutiny. Developers of tools that a jurisdiction has already purchased may decide that they have nothing to gain from letting the defense poke holes in their software.

That explains why even governmental bodies have claimed trade secret protection. The New York City Office of the Chief Medical Examiner has argued for years that the source code for a forensic software program that it developed itself, using public funds, should be privileged. This is absurd: The government has no legitimate commercial interest in keeping details about forensic technology from the defense. Yet the agency has won.

Whether the motive is winning or profit, trade secrets can be abused as a way to keep the defense in the dark. (Last year, a federal judge finally ordered the city to turn over the program’s source code to one defendant; expert witnesses promptly discovered an undisclosed code function likely to aid prosecutors.)

What’s alarming about protecting trade secrets in criminal cases is that it allows private companies to withhold information not from competitors, but from individual defendants.

Even when revealing a company’s techniques could spell economic ruin, there already are ways

to protect them without barring the defense from examining the information. In the business world, companies working on a deal sign non-disclosure agreements promising not to misuse any valuable information that is revealed as part of negotiations. In civil lawsuits, judges often order that proprietary information be shared subject to a protective order, which is like a non-disclosure agreement but with extra sanctions for a violation, such as being held in contempt of court. A federal judge in Delaware once even ordered Coca-Cola to hand over its secret formula in a contract dispute.

The same approach should work in criminal cases.

It is true that the protective order solution can fail. People cheat, especially in a cutthroat industry where only a few people know the technology and the opposing party’s expert witness could be a competitor. As some measure of these anxieties, Coca-Cola chose to concede certain disputed facts rather than comply with the Delaware court’s order.

Even in cases where legitimate business risks exist, though, withholding information from the accused is the wrong answer. Where a company’s business strategy falls on the secrecy-transparency spectrum should not limit the full array of arguments available to a criminal defendant. The law *already* lets businesses sue if someone steals their trade secret. While that might not be a fool-proof solution, making up for any imperfections on the backs of the accused is unfair.

The Supreme Court has an opportunity in *Loomis v. Wisconsin* to rule that the Constitution forbids the government from taking life or liberty based on proprietary secrets. If the court declines, state legislatures should lead the way



by passing laws that direct criminal courts to safeguard valid trade secrets with a protective order and nothing more. It's time to make clear that no one owns the means of decisionmaking in the criminal justice system.

It took weeks for me to get Glenn Rodríguez on the phone. Coordinating collect calls, dialing restrictions, and scheduling from a maximum-security prison takes work. A few times, we

narrowly missed each other. Finally...another colleague emailed to say he was on the line and I dashed into her office to speak with him.

No one had granted his COMPAS reassessment, he said. He had gone in front of the parole board again in January 2017 with the same high risk score.

"I went in there feeling confident about my accomplishments," he told me. "I said, even though the score is high for my prison misconduct, I know that score doesn't represent who I am. I was a different person now, and that would shine through."

The hearing started out like a trial, with the commissioners focused in uncomfortable detail on the crime he had committed a quarter century before. Then, Rodríguez recalled, it shifted.

"You're 43," one commissioner said. "You were 16. You're still young. You still have the opportunity to rebuild your life. We'd like to give you that opportunity."

Rodríguez made parole. He would leave Eastern with 110 college credits from Bard University and a plan to finish his degree and go to graduate school in social work or public health. I asked him if he had any final words on his experience with COMPAS.

"Guys are seeing that pretty much one question can skew the whole thing," he said. "Why is it that there's so much secrecy surrounding this? This is evidence that's being used against you. They are making a determination on a person's life on the basis of this evidence. So you should have a right to challenge it." ■

## About the Author

**Rebecca Wexler** works on data, technology, and criminal justice. She has been a Yale Public Interest Fellow at the Legal Aid Society's criminal defense practice, a Lawyer-in-Residence at the Data and Society Research Institute, and a Justice Stevens Fellow at the Electronic Frontier Foundation. She will begin serving as an assistant professor of law at Berkeley Law School in 2019.



# STATISTICS TEACHER

SUPPORTING THE TEACHING AND LEARNING OF STATISTICS



## THE NEW ONLINE JOURNAL FOR K-12 TEACHERS

**STATISTICS TEACHER (ST)** is an online journal published by the American Statistical Association/ National Council of Teachers of Mathematics Joint Committee on Curriculum in Statistics and Probability for Grades K-12.

- ◆ Timely Articles
- ◆ Peer-Reviewed Lessons Plans
- ◆ Columns
- ◆ News and Announcements
- ◆ Education Publications
- ◆ **And More!**



[www.statisticsteacher.org](http://www.statisticsteacher.org)



# HELP US RECRUIT THE **NEXT GENERATION** OF STATISTICIANS

The field of statistics is growing fast. Jobs are plentiful, opportunities are exciting, and salaries are high. So what's keeping more kids from entering the field?

Many just don't know about statistics. But the ASA is working to change that, and here's how you can help:

- Send your students to [www.ThisIsStatistics.org](http://www.ThisIsStatistics.org) and use its resources in your classroom. It's all about the profession of statistics.
- Download a handout for your students about careers in statistics at [www.ThisIsStatistics.org/educators](http://www.ThisIsStatistics.org/educators).



If you're on social media, connect with us at [www.Facebook.com/ThisIsStats](http://www.Facebook.com/ThisIsStats) and



[www.Twitter.com/ThisIsStats](http://www.Twitter.com/ThisIsStats). Encourage your students to connect with us, as well.

## Site features:

- Videos of young statisticians passionate about their work
- A myth-busting quiz about statistics
- Photos of cool careers in statistics, like a NASA biostatistician and a wildlife statistician
- Colorful graphics displaying salary and job growth data
- A blog about jobs in statistics and data science
- An interactive map of places that employ statisticians in the U.S.