ARTICLES

"Let 'em Play"
A Study in the Jurisprudence of Sport

MITCHELL N. BERMAN*

TABLE OF CONTENTS

---

* Richard Dale Endowed Chair in Law, Professor of Philosophy (by courtesy), The University of Texas at Austin. Earlier drafts of this Article were presented at the 2010 Analytical Legal Philosophy Conference at NYU, the 2010 Annual Meeting of the International Association for the Philosophy of Sport held at the University of Rome, and at faculty workshops at the University of Texas School of Law and the McMaster University Philosophy Department. I am grateful to participants at these events for their reactions and criticisms, to Ronen Avraham, Aaron Bruhl, David Enoch, Rich Friedman, Marshall King, Matt Kramer, John Russell, Stefan Sciaraffa, Fred Schauer, Seana Shiffrin, Matt Spitzer, Abe Wickelgren, and Ben Zipursky for very helpful written comments, and to Rich Friedman (again) for the title. Shane Anderson, Anthony Arguijo, and Jamie Leader provided excellent research assistance.

INTRODUCTION: ONE STEP OVER THE LINE

Kim Clijsters was the feel-good story of the 2009 U.S. Open. A former world #1, the twenty-six-year-old Belgian had been retired for two years, during which time she had married and borne a child, when she surprised the tennis world by announcing her return in the summer of 2009.[1] Entering the tournament as an unranked wildcard, Clijsters defeated Denmark's Caroline Wozniacki in straight sets to become the first unseeded player ever to win the U.S. Open and the first mother to win a grand slam event in thirty years.[2]

The match that made the headlines, however, was not the final. It was Clijsters's semifinal contest against the #2 seed, Serena Williams. Straight off her victory at Wimbledon, Williams had powered through the women's draw at the Open without losing a set; with the third-round defeat of #1 seed Dinara Safina, she was the odds-on favorite to win her third grand slam tournament of the year. Instead, Williams lost the first set 4–6 and found herself serving to Clijsters at 5–6 in the second. Down 15–30, Williams's first serve to the advantage court was wide. On her second serve, the line judge called Williams for a foot fault, putting her down double-match point. At this, Williams exploded, walking over to the judge several times, gesticulating with her racket in a menacing manner, shouting, and threatening to do things with the ball that the lineswoman was bound to find unwelcome. Because Williams had already committed a code violation earlier in the match for racket abuse, this second code violation called forth a mandatory one-point penalty—that single point gave the match to Clijsters.[3]

Williams has few defenders. Her outburst was further penalized with the maximum onsite penalty of $10,000 and later an additional $82,500 and two-

1. *See Tennis Star Kim Clijsters Announces Retirement*, FOXNEWS.COM (May 6, 2007), http://www.foxnews.com/story/0,2933,270274,00.html; *US Open 2009: Kim Clijsters Makes Winning Return to New York*, THE TELEGRAPH (Aug. 21, 2009), http://www.telegraph.co.uk/sport/tennis/usopen/6118406/US-Open-2009-Kim-Clijsters-makes-winning-return-to-New-York.html.

2. Tina Molly Lang, *Kim Clijsters Defeats Caroline Wozniacki to Win US Open*, ASSOCIATED CONTENT (Sept. 14, 2009), http://www.associatedcontent.com/article/2172747/kim_clijsters_defeats_caroline_wozniacki.html; *Unseeded Clijsters Secures U.S. Open Victory*, CNN.COM (Sept. 14, 2009), http://www.cnn.org/2009/SPORT/09/13/tennis.us.open.womens/index.html.

3. For a detailed description of the match, see John Martin, *In a Bizarre Ending, Clijsters Beats Williams*, STRAIGHT SETS (Sept. 12, 2009), http://straightsets.blogs.nytimes.com/2009/09/12/live-analysis-wickmayer-vs-wozniacki/.

year probation.[4] But without condoning or excusing Williams's response to the foot-fault call, I'm interested in a different question: whether the call should have been made at all. CBS color commentator and former tennis great John McEnroe thought not. As he remarked at the time: "You can't call that there."[5] His point was not that the call was factually mistaken,[6] but rather that, even assuming arguendo that it was factually supportable, it was an inappropriate call to make at that point in the match: the lineswoman should have cut Williams a little slack. Many observers agreed. As another former tour professional put it, a foot fault is "something you just don't call—not at that juncture of the match."[7]

The McEnrovian position—that at least some rules of some sports should be enforced less strictly toward the end of close matches—is an endorsement of what might be termed "temporal variance." It is highly controversial. Indeed, some people find it simply incredible that a call conceded to be correct at one time could be thought improper at another. As one letter writer to the *New York Times* objected: "To suggest that an official not call a penalty just because it happens during a critical point in a contest would be considered absurd in any sport. Tennis should be no exception."[8] On this view, which possibly resonates with a common understanding of what it means to follow "the rule of law," rules of sports should be enforced with resolute temporal invariance.

Now, perhaps McEnroe was wrong about the Williams foot fault. But the premise of the letter propounding the competing view—that participants and fans of any other sport would reject temporal variance decisively—is demonstrably false. Indeed, one letter appearing in *Sports Illustrated* objected to the disparity of attention focused on Williams as compared to U.S. Open officials, precisely on the grounds that "[r]eferees for the NFL, NHL and NBA have generally agreed that in the final moments, games should be won or lost by the players and not the officials."[9] I am unsure just how general this supposed agreement is. But I'd warrant that most fans of professional basketball would affirm that contact that would constitute a foul through most of the game is frequently not called during the critical last few possessions of a close contest.

---

4. *Serena Williams Fined Record $82,500 for U.S. Open Outburst*, USA TODAY (Dec. 1, 2009), http://www.usatoday.com/sports/tennis/2009-11-30-serena-williams-open-fine_N.htm.

5. *See* Geoff MacDonald, *The Thinking Behind Calling Foot Faults*, STRAIGHT SETS (Sept. 13, 2009), http://straightsets.blogs.nytimes.com/2009/09/13/the-thinking-behind-calling-foot-faults/.

6. The rule governing foot faults provides that, "During the service motion, the server shall not . . . [t]ouch the baseline or the court with either foot . . . ." INT'L TENNIS FED'N, RULES OF TENNIS 2010, at 8 (2009) (Rule 18.b); *see also* U.S. TENNIS ASS'N, FRIEND AT COURT 12 (2010) ("Comment 18.3: *When does a foot fault occur?* A player commits a foot fault if after the player's feet are at rest but before the player strikes the ball, either foot touches . . . the court, including the baseline . . . .").

7. Michael Wilbon, *A Call and a Response that Can't Be Defended*, WASH. POST, Sept. 14, 2009, at D3, *available at* http://www.washingtonpost.com/wp-dyn/content/article/2009/09/13/AR2009091302533.html.

8. Vince Bray, Letter to the Editor, N.Y. TIMES, Sept. 20, 2009, at 5, *available at* http://query.nytimes.com/gst/fullpage.html?res=9B01E6DB1439F933A1575AC0A96F9C8B63.

9. J. Everett Prewitt, Letter to the Editor, SPORTS ILLUSTRATED (Oct. 12, 2009), http://sportsillustrated.cnn.com/vault/article/magazine/MAG1161011/index.htm.

Moreover, most fans I have spoken with believe this is not only how it is, but how it ought to be.[10] In any event, precious few of those who disagree would contend that the status quo is absurd. So an insistence on rigid temporal invariance requires argument, not just assertion.

However, advocates of temporal variance ought not to be too smug either. For while the negative import of temporal variance is clear—namely, the *denial* of categorical temporal *in*variance—its positive import is anything but. Surely those who believe that the foot fault ought not to have been called against Serena Williams in her match against Kim Clijsters mean implicitly to invoke a principle broader than "don't call foot faults in the twelfth game of the second set of semifinal matches in grand slam tournaments." But how much broader? Is the governing principle that *all* rules of *all* sports should be enforced less rigorously toward the end of contests? Presumably not. Few proponents of temporal variance, I'd wager, would contend that pitchers should be awarded an extra inch or so around the plate in the ninth inning, or that a last-second touchdown pass should be called good if the receiver was only a little bit out of bounds. So even if categorical temporal invariance is too rigid, the contours and bases of optimal temporal variance remain to be argued for. That is the task of this Article. Or, I should rather say, that is this Article's surface agenda.

While seeking a deeper understanding of familiar sporting practices is a worthwhile project all on its own, this investigation is in service of a greater ambition. Jurisprudes have long drawn on games and sports for illumination.[11] Rawls, for example, famously drew on baseball to exemplify his practice conception of rules.[12] Hart invoked chess and cricket to illustrate both the internal aspect of rules and the difference between primary and secondary

10. An anonymous reader of this Article expressed puzzlement that I should appeal to intuitions widely held by sports fans, opining that "the point of introducing the analytical apparatus" should instead be "to provide a structure within which those intuitions could be corrected." Well, yes, of course, the analytical apparatus will allow us to correct intuitions when they should be corrected. But that we should not hold ourselves hostage to pre-reflective intuitions does not mean that we should disregard them entirely. Reflective equilibrium—the method that I employ—insists that our general principles and "analytical apparat[i]" do not enjoy categorical priority over our case-specific judgments. Rather, it counsels that we try to bring into coherence our considered judgments about principles at various levels of generality and case-specific outcomes.

11. Just what games and sports are, and whether the latter is a proper subclass of the former, are questions that I do not explore in this Article. Briefly, the Wittgensteinian teaching that games cannot be defined in terms of necessary and sufficient conditions was challenged some decades ago in a short, little-noticed book—BERNARD SUITS, THE GRASSHOPPER: GAMES, LIFE AND UTOPIA (Broadview Encore 2005) (1978)—that has recently won praise from such philosophers as Thomas Hurka and Simon Blackburn, *see id.* at 7–20 (introduction by Thomas Hurka); *Simon Blackburn Defends the Grasshopper*, VIRTUAL PHILOSOPHER (Feb. 6, 2008), http://nigelwarburton.typepad.com/virtualphilosopher/2008/02/simon-blackburn.html. I am uncertain that Suits's effort succeeds. For a criticism, see Norman Geras, *Games and Meaning*, *in* HILLEL STEINER AND THE ANATOMY OF JUSTICE 185 (Stephen de Wijze et al. eds., 2009). Be that as it may, I am disposed to think John Tasioulas right, as against Hurka and Suits, in insisting that not all sports are games. *See* Thomas Hurka & John Tasioulas, *Games and the Good*, ARISTOTELIAN SOC'Y, July 2006, at 217.

12. *See* John Rawls, *Two Concepts of Rules*, 64 PHIL. REV. 3, 25 (1955).

rules.[13] Dworkin also turned to chess to demonstrate the notion of constructive interpretation.[14] Examples like these could be multiplied with ease. Yet, jurisprudential attention to sports and games is decidedly ad hoc. I am unaware of any sustained or systematic investigation into the insights that formal sports and municipal legal systems might offer up for students of the other.[15]

The lack of sustained jurisprudential attention to games and, especially, sports should surprise, for sports leagues plainly constitute distinct legal systems. This should be apparent on the surface—at least to non-Americans. While the American sports scene is dominated by three home-grown team sports—baseball, football, and basketball—all of which are governed by official "rule books," the most popular global team sports like soccer, cricket, and rugby (both league and union)[16] are all formally governed by "laws," not "rules."[17] But the law-ness of sports systems is not merely superficial, for they exhibit such essential institutional features as legislatures, adjudicators, and the union of primary and secondary rules.[18] We might reasonably have expected a discipline of "sports and law" to have arisen as a region of study belonging either to comparative law or to special jurisprudence.

To be sure, philosophy has spawned a subfield-denominated philosophy of sport, actively represented by at least two English-language societies (the International Association for the Philosophy of Sport and the British Philosophy

13. *See* H.L.A. HART, THE CONCEPT OF LAW 89 (2d ed. 1994).

14. *See* RONALD DWORKIN, LAW'S EMPIRE 136–38 (1986); *see also* RONALD DWORKIN, TAKING RIGHTS SERIOUSLY 101–05 (1977).

15. Indeed, most legal writing on sports that does not pertain to sports law is intended more to entertain than to edify. The best known example of the genre is Aside, *The Common Law Origins of the Infield Fly Rule*, 123 U. PA. L. REV. 1474 (1975). Recently, however, John Roberts's proposed analogy, offered during his confirmation hearings, between judging and umpiring, has provoked interesting jurisprudential writing—most of it (rightly) critical. *See, e.g.*, Aaron S.J. Zelinsky, *The Justice as Commissioner: Benching the Judge–Umpire Analogy*, 119 YALE L.J. ONLINE 113 (2010); Neil S. Siegel, *Umpires at Bat: On Integration and Legitimation*, 24 CONST. COMMENT. 701 (2007). As Judge Richard Posner objected, "Neither [Roberts] nor any other knowledgeable person actually believed or believes that the rules that judges in our system apply, particularly appellate judges and most particularly the Justices of the U.S. Supreme Court, are given to them the way the rules of baseball are given to umpires." RICHARD A. POSNER, HOW JUDGES THINK 78 (2008).

16. Amrit Doley, *The World's Top 10 Most Popular Team Sports*, SPORTINGO (May 16, 2009), http://www.sportingo.com/all-sports/a11587_worlds-top-most-popular-team-sports.

17. Cricket is governed by the "Laws of Cricket," *see Laws of Cricket*, LORD'S.ORG, http://www.lords.org/laws-and-spirit/laws-of-cricket/laws/ (last visited Jan. 22, 2011), soccer by the Fédération Internationale de Football Association (FIFA) Statutes, FÉDÉRATION INTERNATIONALE DE FOOTBALL ASS'N, FIFA STATUTES: AUGUST 2009 EDITION (2009), and rugby by the "Laws of the Game Rugby Union" and "Laws of Rugby League," INT'L RUGBY BD., LAWS OF THE GAME RUGBY UNION (2007); RUGBY FOOTBALL LEAGUE, THE INTERNATIONAL LAWS OF THE GAME AND NOTES ON THE LAWS (2004).

18. *See About FIFA*, FIFA.COM, http://www.fifa.com/aboutfifa/index.html (last visited Jan. 22, 2011); *About ICC*, INT'L. CRICKET COUNCIL, http://icc-cricket.yahoo.net/the-icc/about_the_organisation/overview.php (last visited Jan. 22, 2011); *IRB Organisation*, INT'L RUGBY BD., http://www.irb.com/aboutirb/organisation/index.html (last visited Jan. 22, 2011); *Operational Rules*, RUGBY FOOTBALL LEAGUE, http://www.therfl.co.uk/operational_rules (last visited Jan. 22, 2011).

of Sport Association)[19] that publish two specialty journals (*Journal of the Philosophy of Sport* and *Sport, Ethics, and Philosophy*). But the coherence of this discipline is unclear. Its ecumenical embrace of nearly the entire panoply of standard philosophical subfields—one resource proclaims that it draws on aesthetics, epistemology, ethics, logic, metaphysics, philosophy of education, philosophy of law, philosophy of mind, philosophy of rules, philosophy of science, and social and political philosophy[20]—invites a prediction that the philosophy of sport might not long survive as a unified domain of philosophical study.[21] Much more to the present point, even insofar as philosophy of sport encompasses the philosophy of law and "the philosophy of rules," it has not apparently drawn interest from scholars actively engaged in legal philosophy.

I think that is as unfortunate as it is surprising. Legal theorists, after all, might be thought to be the experts in the problematics of rule-governed social institutions. The grander ambition of this Article, accordingly, is to help spur the growth of the jurisprudence of sport, or of the philosophy of sports and law, as fields worthy of more concerted theoretical attention. One might even say that this Article does double duty as a manifesto of sorts for an enlarged program of jurisprudential inquiry.

Because this goal must strike some readers, and not only the more sports-phobic, as quixotic, perhaps I should note just some of the ways in which, as formal rule-governed practices, sports and law often pursue similar goals and confront many of the same challenges. For example, each domain must decide: to what extent to guide conduct by "formal" written norms as opposed to "informal" social norms, and, if the former, by rules or by standards; when to delegate discretion to the adjudicators (judges, juries, referees), and how best to constrain that discretion; how to respond to the problem of epistemic uncertainty; whether to provide a right of appeal from unfavorable decisions and, if so, how to structure appellate review; how to conceptualize, deter, and sanction "cheating"; how to identify and rectify the gaps that inevitably arise between "the law in the books" and "the law in action"; when to tolerate ties and how to resolve them when they should not be tolerated; how to analyze and craft optimal sanctions; and so on and so forth.

Finally, it's not just that (municipal) legal systems and sports systems confront similar challenges. There are reasons to believe that jurisprudential atten-

---

19. BRITISH PHIL. SPORT ASS'N, http://www.philosophyofsport.org.uk/ (last visited Jan. 22, 2011); INT'L ASS'N PHIL. SPORT, http://www.iaps.net (last visited Jan. 22, 2011).

20. LEON CULBERTSON ET AL., HIGHER EDUC. ACAD., RESOURCE GUIDE TO THE PHILOSOPHY OF SPORT AND ETHICS OF SPORT 1 (2008), http://www.heacademy.ac.uk/assets/hlst/documents/resources/philosophy_ethics_sport.pdf.

21. *Cf.* Frank H. Easterbrook, *Cyberspace and the Law of the Horse*, 1996 U. CHI. LEGAL F. 207, 207–08 ("[T]he best way to learn the law applicable to specialized endeavors is to study general rules. Lots of cases deal with sales of horses; others deal with people kicked by horses; still more deal with the licensing and racing of horses, or with the care veterinarians give to horses, or with prizes at horse shows. Any effort to collect these strands into a course on 'The Law of the Horse' is doomed to be shallow and to miss unifying principles.").

tion to the sporting domain is particularly likely to contribute to our understanding of the phenomena and dynamics shared in common. First, because the rules and practices of sports have long been viewed as unworthy of serious philosophical and theoretical investigation, even low-hanging fruit has yet to be harvested. Second, as we will see, sports supply a vast range of examples for the generation of hypotheses and against which to test our theories. And third, our judgments and intuitions about certain practices—such as, to take the present topic, the propriety of (one type of) context-variant enforcement of rules—are less likely in the sports courts than in the courts of law to be colored or tainted by possibly distracting substantive value commitments and preferences.

In short, sporting systems, though rarely explored with seriousness by legal theorists and comparative lawyers, comprise a worthy object of legal–theoretical study. This Article is offered as an illustration of that worth.

I. PRELIMINARIES AND THE PLAN OF ATTACK

My broader ambition notwithstanding, our narrow subject remains whether, and under what circumstances, rules of sports should be enforced with greater laxity toward the end of close contests. Before tackling that question, let us consider a few caveats.

First, I mean this "should" in a legal, not a moral, sense. We have noted the possibility that the lineswoman should have called a foot fault on Serena Williams just so long as she genuinely believed that Williams's foot touched or went over the baseline before hitting the ball. Those who think otherwise could reach the contrary conclusion via varied routes. One possibility is to go outside the legal system. A proponent of this approach would agree that the rules of tennis concerning foot faults should be crafted and understood to be temporally invariant but would also argue that "moral" or "all-things-considered" reasons dictated that the lineswoman should have refused to enforce the rule nonetheless. In directing our attention to the legal "should," I mean to put that possibility aside. I am interested in the question of when the legal regime that makes up tennis should itself provide for temporally variant enforcement of foot faults.

Second, I reject the wholly conventionalist answer to this question. Some people think that what I claim to be a puzzle of genuine legal–philosophical interest is no puzzle at all. They say that the answer is that it all depends on the "norms," "customs," or "conventions" of the sport in question—a matter about which they may cheerfully admit to have no expert knowledge. But this won't do. For we are seeking not simply a report of existing practices but an account of what the practices should be. If, for example, it is the custom in some sport to afford competitors greater slack in certain game situations, the proponents of that custom should believe that it is backed by good reasons and should hope to grasp what those good reasons are.

This is not to say that sport-specific norms and customs are irrelevant. Golf

has a very different internal morality or ethos than does baseball.[22] It might well be that these differences properly bear on optimal practices regarding temporal variance. All I mean is that we should not assume that the existing practices concerning temporal variance are conclusive regarding what those practices should be, all things considered, even after we adequately account for relevant sport-specific features.

Third, when investigating the legal "should," I intend at this early stage of the analysis to be agnostic among the variety of ways that an affirmative answer could be operationalized. Here, for instance, are five ways that temporal variance for foot faults could be realized: (1) the rules should be drafted expressly to specify the contexts in which touching the line or court does not constitute a foot fault; (2) the rules should be drafted expressly to specify the contexts in which touching the line or court, while an infraction, is unenforceable; (3) the rules should be drafted in a manner that does not expressly rule out context variance and should be interpreted and enforced in a context-variant manner; (4) the rules should be drafted and interpreted to confer discretion on the linesperson or umpire to adjudge infractions in a context-variant manner; or (5) the rules should be drafted and interpreted in a manner that denies context variance to adjudicating the fact of infraction but confers discretion on the linesperson or umpire whether to enforce the prescribed penalty for the infraction. I am not at the outset distinguishing among these (or other) routes to temporal variance. Rather, the question of whether foot faults should be enforced in a temporally variant manner is essentially whether any one of these (or other roughly similar) propositions is true. I interpret temporal invariance to be committed to the proposition that none of these answers is true. However, if we conclude in favor of temporal variance, we can then explore whether the arguments in favor of temporal variance also bias in favor of a particular way to implement it. Put slightly differently, assuming arguendo that foot fault penalties should be enforced with temporal variance (and therefore that, ideally, Williams should

22. More noteworthy than golf's requirement that participants call infractions on themselves is the fact that players routinely do—even when the infraction was witnessed by no one else and when it conferred no competitive benefit. *See, e.g.*, Tom Yantz, *Take a Bow Brian Davis for Calling Infraction on Himself*, ON THE TEE (Apr. 19, 2010), http://blogs.courant.com/chip_shots/2010/04/take-a-bow-brian-davis-for-cal.html. The legendary Bobby Jones, competing in the 1925 U.S. Open, conveyed the extent to which golfers have internalized their sport's code of honor when he assessed himself a one-stroke penalty after he alone saw his ball move a fraction of an inch when he addressed it in the rough. ROBERT SOMMERS & ARNOLD PALMER, GOLF ANECDOTES: FROM THE LINKS OF SCOTLAND TO TIGER WOODS 81–82 (2004). Applauded afterwards for his integrity, Jones would have none of it: "You might as well praise me for not robbing banks," he protested. *Id.* at 82. Baseball, in contrast, is widely understood, even glorified, as a game of cheating and deception—from spitballs and sign stealing to the hidden-ball trick. *See generally* JASON TURBOW & MICHAEL DUCA, THE BASEBALL CODES (2010). As Chicago Cubs President Andy MacPhail noted when his star slugger, Sammy Sosa, was caught with a corked bat, "There is a culture of deception in this game. It's been in this game for 100 years. I do not look at this in terms of ethics. It's the culture of the game." *Id.* at 184. Of course, even baseball does not tolerate all deception, though the precise location and contours of the line separating the permissible from the impermissible is contested.

not have been penalized), I am not presently focused on which agent of the tennis system—the gamewrights, the line judge, the chair umpire, etc.—should have done differently than they, she, or he did.

Fourth, my analysis is *pro tanto*, not conclusive. Figuring out whether a practice of temporal variance is optimal for any given sport, all things considered, is devilishly hard—too many considerations enter the scene, implicating too many contestable empirical and evaluative premises. But perhaps at this early stage we should content ourselves with the more modest ambition of trying to figure out whether "sense can be made" of such a practice. Instead of trying to furnish a conclusive argument about what the optimal practices should be, I will consider my efforts a success to the extent that this Article explains why temporally variant rule enforcement might be sensible—or what can plausibly be said for such a practice. Because many readers appear to forget this proviso by Article's end, I reiterate: this Article does not claim to present an all-things-considered argument for temporal variance.

So much for preliminaries. Here is a short overview of the remainder of this Article.

Part II begins, not with tennis, but with other sports in which it has been alleged that a practice of temporal variance is more secure—sports like football, hockey, and basketball. In each, whistles for minor physical contact toward the end of tight contests predictably elicit a cry from the stands: "Let 'em play!" Though the plea is familiar, its rationale is not obvious. To be sure, the tighter the rules are enforced, the less physical contact there will be. And reasonable fans and participants may reasonably disagree about the level of physicality that makes the sport in question the best it can be. But, however we—or the respective leagues—may answer that question, it is not self-evident why the optimal degree of laxity should be any different in crunch time during an NBA game, or throughout the NHL playoffs, than at any other time. In other words, it is not obvious what can be said for "letting them play" *at this particular time* different in character or force from what can be said *generally* for "letting them play."

No, it is not obvious. Still, this is a good place to start. I'm skeptical that many tennis fans could assert with confidence that tennis officials allow players to get away with more foot faults toward the end of close matches than earlier. Maybe they do, but foot faults just aren't called enough at any time to permit those without intimate knowledge of the sport to be sure what the patterns of enforcement are.[23] Things are different with basketball. That basketball referees

23. *See, e.g.*, Mark Sappenfield, *Serena Williams Foot Fault: What Did She Say and Why?*, CHRISTIAN SCI. MONITOR (Sept. 13, 2009), http://www.csmonitor.com/USA/2009/0913/p02s01-usgn.html ("[T]he [foot-fault] rule is . . . inconsistently enforced. A player must not touch any part of the service line during a serve, yet line judges often ignore infractions. . . . [S]ome players say there is an unwritten rule that—just as hockey referees call nothing but the most blatant penalties in overtime playoff games—tennis officials should ignore seemingly ticky-tack infractions like foot faults when stakes are high.").

respect some measure of temporal variance seems clear enough to many hoops fans.[24] Maybe that is because the case for temporal variance in basketball is unusually clear. If we can explain and justify slack in the calling of basketball fouls, we might be in a stronger position to assess whether temporal variance makes sense in tennis too. Unfortunately, whether the analysis that I end up offering in support of temporal variance in basketball and football applies to foot faults is uncertain. Part III explains why and proposes an alternative analysis that can help explain temporal variance in that context too. Part IV offers brief concluding thoughts and discusses a surprising implication for improving the world's most popular sport.

II. SWALLOWING THE WHISTLE: TEMPORAL VARIANCE IN THE ENFORCEMENT OF FOULS

I'm with those who believe that fouls are frequently called less strictly at the end of close games than otherwise. Even if true, this is hard to establish to the skeptical, for if this is the rule, it is an example of "the rule in action," not "the rule in the books." Those portions of the official NBA rules that define and proscribe personal fouls do not provide that the kind or amount of contact required to constitute a foul varies according to contest-contextual features.[25]

The most general statement of personal fouls provides that "[a] player shall not hold, push, charge into, impede the progress of an opponent by extending a hand, forearm, leg or knee."[26] More particularly, "[c]ontact initiated by the defensive player guarding a player with the ball is not legal," although "[i]ncidental contact with the hand against an offensive player shall be ignored if it does not affect the player's speed, quickness, balance and/or rhythm."[27]

Contrast this formal invariance with the NBA's remarkable varying standard of proof. As a comment on the administration and application of the rules directs, "there are times during a game where 'degrees of certainty' are necessary to determine a foul during physical contact. This practice may be necessary throughout the game with a higher degree implemented during impact times

24. In the first round of the 2010 NCAA men's basketball tournament, a referee called a lane violation against a player for New Mexico State with 18.6 seconds remaining and the New Mexico State Aggies down to Michigan State 69–67. *See Lucas, Morgan Save MSU in 70–67 Win Over NMSU*, SPORTS ILLUSTRATED (Mar. 19, 2010), http://sportsillustrated.cnn.com/basketball/ncaa/men/gameflash/2010/03/19/62933_recap.html. The call gave the Spartans another free throw, which they made for a three-point lead that they didn't relinquish. *Id*. Commentators were withering in their criticism of the call. "That's a horrendous, horrendous call," objected one. "Lane violations happen on a majority of free throws and are almost never whistled. To call one with 18 seconds left in a two-point NCAA tournament game is unconscionable." Steve Schrader, *Critics: Refs Blew It on Lane Violation*, DETROIT FREE PRESS, Mar. 21, 2010, at C8. Importantly, the objection was not merely that the call was made inconsistently. Rather, protested one observer, "You can't have that be a deciding call in a close NCAA tournament game." *Id.*

25. *See* NAT'L BASKETBALL ASS'N, OFFICIAL RULES 2006–2007, at 43 (2006) (Rule 12(B)(I)).

26. *Id.*

27. *Id*.

when the intensity is risen, especially nearing the end of a game."[28] Notice, then, that the formal rules that specify what *constitutes* a foul are context-invariant, but the standards of proof that determine whether a particular action will be *adjudged* to be of the forbidden type are context-variant.

An example might help make the contrast clearer. Suppose Referee is very confident that a defensive player, say the Spurs' Tony Parker, makes incidental contact with his hand against an offensive player, say the Miami Heat's Dwyane Wade. But Referee is very uncertain whether that contact affected Wade's speed, quickness, balance, or rhythm. The "degree of certainty" rule seems to dictate something like the following: if Referee believes that it is only modestly more likely than not that Parker's contact affected Wade's speed, (1) he should call a foul through most of the game; and (2) he should not call a foul with the game tied in the last minute. In contrast, as far as the formal rules provide, were Referee fairly confident that Parker's contact did affect Wade's speed (by slowing him), but only to a slight degree, then he should enforce the foul no matter when it occurred. One way to think about our task is as an attempt to figure out what could be said for bringing greater symmetry to the rules—for allowing the referee the same discretion in determining how much contact (or how much impedance) constitutes a foul at different points during the game as he is granted in determining how confident he must be that the forbidden amount of contact was committed.

One answer invokes essentially aesthetic considerations: the referee's whistle disrupts play, thereby reducing spectators' enjoyment of the action. Disruption of play almost always incurs an aesthetic cost. But inasmuch as dramatic tension builds during crunch time, disruption of play during this time is especially costly.

There is something to this justification for temporal variance. It would seem to apply, though, only when play would continue uninterrupted but for the calling of a foul. However, in some sports that arguably respect temporal variance, play stops either way. It appears to me, for example, that football officials are often a little more reluctant to call defensive pass interference during crunch time even though an incompletion stops play just as surely as does a penalty flag.[29] Because an aesthetic or dramatic preference that play not

---

28. *Id*. at 51 (Comments on the Rules: I. Guides for Administration and Application of the Rules).

29. And not only to me. Here is *Sports Illustrated*'s lead NFL columnist, Peter King, explaining his naming Browns cornerback Hank Poteat "goat of the week":

> [W]ith Cleveland holding a 37–31 lead and no time left on the clock in the fourth quarter, Detroit quarterback Stafford let fly with a rainbow to the end zone and Poteat tackled Calvin Johnson with the ball in the air. *If Poteat had jostled Johnson, there's little chance a flag would have been thrown*. But a full-scale body slam to the ground . . . That has to be called. Pass interference. With the extra play, Detroit threw a touchdown pass to win it. On the goat scale, Poteat's play ranks about as high as you can go.

Peter King, *Monday Morning QB*, SPORTS ILLUSTRATED (Nov. 23, 2009), http://sportsillustrated.cnn.com/2009/writers/peter_king/11/22/Week11/3.html (omission in original) (emphasis added).

be disrupted would seem not to explain or justify temporal variance in the calling of fouls and the enforcement of penalties across the board, it might not provide the whole story even in basketball. So without denying that appreciation for dramatic excitement can help explain why officials should give the competitors somewhat greater slack during moments of high drama, it behooves us to explore the possibility of an alternative account.

The alternative account that I'll offer has two core components. First, expanding on the wisdom at the heart of the saying "no harm, no foul," I argue in sections II.A through II.C that, *insofar as penalties are designed to serve a compensatory or restitutionary function*, we have reason not to impose them when they would work substantial overcompensation.[30] Second, starting from the straightforward idea that the competitive cost imposed by any given outcome-affecting contest event[31] is not constant, but context-variant, I draw attention to one contextual factor of particular significance: time (or functional equivalents). In particular, I argue in sections II.D and II.E that the competitive impact of an event occurring during a close contest is inversely proportional to the distance remaining to the contest's completion.[32] If these two claims are correct, then it follows that a penalty of nominally constant magnitude optimal to impose early in a contest may become suboptimal to impose later in that same contest.

A. ON THE SAYING, "NO HARM, NO FOUL"

Take a step back and consider the familiar saying "no harm, no foul" (NHNF). We hear and use it frequently. But what exactly does it mean?

I'm not the first to ask this question. "Curious" posed it a few years ago on *Yahoo! Answers*, eliciting this "best response": "No one was hurt, nothing is wrong."[33] Another reply offered, "it means that if [no one was] hurt in what you

30. The italicized qualification is significant for at least two reasons. As we will see, penalties that have a restitutionary function—that is, penalties that serve to offset a competitive advantage that a competitor has gained from violating a rule—almost invariably have a deterrent function too. Moreover, some penalties have no compensatory or restitutionary function at all—as when the rules to which they attach serve not to police a competitive balance but rather to prohibit, for example, dangerous or unsporting behavior. Penalties that attach to such rules principally serve deterrence and possibly retribution too, but not compensation or restitution.

31. Outcome-affecting contest events include scores, infractions of the rules that confer competitive benefits on the rule breakers, penalties imposed by officials in response to infractions, etc.

32. Many sports (for example, basketball, football, hockey, soccer) are timed. In such sports, distance to contest completion is temporal. Unclocked sports (for example, baseball, tennis, golf, volleyball) use mechanisms other than the passage of time to determine when a contest ends. In those sports, distance to completion must be measured in other units. In baseball, the units are outs that a team may incur or must secure; in tennis and volleyball, they are games that one must win or lose; in golf, the units are holes. So these other means of measuring progress toward contest completion—outs, games, holes, etc.—are what I mean by "functional equivalents" of time. For simplicity of exposition, I will speak about time and temporality but with the understanding that I mean to invoke its functional equivalents when the context requires.

33. *What Is the Exact Meaning of No Harm No Foul?*, YAHOO! ANSWERS, http://answers.yahoo.com/question/index?qid=20070518143109AAx6BR1 (last visited Apr. 29, 2011).

were doing, then what you have done is justified."[34] A second website, *UsingEnglish.com*, elaborates: "There's no problem when no harm or damage is done, such as the time my sister-in-law stole the name we'd chosen for a boy and we both ended up having girls."[35]

This is bad ethics. Take the case of the thieving sister-in-law. Admittedly, I'm not entirely certain what is involved in stealing a name; despite what a surprising number of parents appear to believe, one *can* give a child a name already in use.[36] Still, I imagine the scenario goes like this: One pregnant woman learns of the name that her pregnant sister (or sister-in-law) intends to give her baby if it is a boy. She then forms the intention to give it to her own baby, if a son, knowing that doing so would make the name no longer attractive, or significantly less attractive, to the couple who had the idea before her. That both women bore girls softens the sting of the betrayal, but the notion that the sister-in-law did nothing wrong, or that "there's no problem" with her conduct, seems plainly mistaken. We are told, after all, that she "stole" something (a name). That would seem to be wrongful if true. That the wrong proved harmless is fortunate, but not an erasure of the wrong itself.

The answers proposed by the *Yahoo!* responders, though admirably concise, are even further off base. One can run afoul of standards of rightful or permissible behavior even without causing harm, which is why almost nobody thinks criminal punishment for unsuccessful complete attempts is unjust or morally impermissible (assuming that criminal punishment is not generally unjust or impermissible). So if "no harm, no foul" has any sensible meaning, we still haven't identified just what it is.

True story: some evenings ago my wife and I returned home to find our house empty, but the front door unlocked. Our kids' babysitter had taken them to the park and had forgotten to lock up. My wife remarked upon it when he returned. "Gosh," he said embarrassedly. "You're right. I just forgot. I'm really sorry." "Well, it's okay," she assured him. "No harm, no foul."

That, I think, is a sound usage, but my wife wasn't saying that our babysitter's lapse was "justified," nor was she denying that he had, in fact, done something wrong—an unintentional wrong, but a wrong nonetheless. At least in my wife's deployment of the phrase, "no harm, no foul" is a performative—namely, an acceptance of an apology. Thus does another web dictionary define the phrase: "[N]o problem, it's cool. Usually used in response to someone's apology to indicate acceptance."[37]

---

34. *Id.*

35. *Idiom Definition: No Harm No Foul*, USINGENGLISH.COM, http://www.usingenglish.com/reference/idioms/no+harm,+no+foul.html (last modified Apr. 27, 2011).

36. Nearly thirty percent of black girls born each year in California receive a name given to no other baby born in the state. STEVEN D. LEVITT & STEPHEN J. DUBNER, FREAKONOMICS: A ROGUE ECONOMIST EXPLORES THE HIDDEN SIDE OF EVERYTHING 184 (2005).

37. Ryan, *No Harm No Foul*, URBAN DICTIONARY (Feb. 22, 2005), http://www.urbandictionary.com/define.php?term=no%20harm%20no%20foul.

Insofar as this is correct, it has an important implication for our question. To see why, we should distinguish two things that are often conflated: accepting an apology and demurring to it.

Sometimes we are causally responsible for bad states of affairs even when not to blame. Take this common example: *D* is driving down a residential street at an appropriate speed and with utmost care. *C*, a young child, darts out from behind a parked car into the path of *D*'s minivan. With no time to react, *D* drives into *C*, killing him. *D* has a relationship to the event that *P*, a pedestrian down the block, lacks: *D* caused the death of a child. Yet, on these facts, he is not to blame. Moreover, on the dominant view, he hasn't even committed a wrong.[38] It is tragic all around—for *C*, *C*'s family, and *D* himself—yet, for all that, it is a quintessentially blameless accident. However, *D*'s blamelessness doesn't let him entirely off the hook. Bernard Williams famously contended that *D* ought to experience agent-regret.[39] At a minimum, he has a duty to express remorse: "I'm so terribly sorry."

It seems to me that *C*'s family, if they understand the relevant basic facts, ought to "let *D* off the hook." There are different ways to do so. (Sadly, in this case, "no harm, no foul" is not among them.) Imagine this choice: "Thank you. We accept your apology." If you're *D*, you might think this not quite apt. You might feel, whether or not you choose to voice it, that by "I'm sorry" you didn't mean "I apologize"—at least not in a robust or unqualified sense. You felt deep regret for the outcome and remorse for your causal role but were not intending to own a wrong. The better response of *C*'s family would have been to acknowledge that fact. "It wasn't your fault" would have been better than "We accept your apology."[40]

The point of this sad little story is that accepting an apology is not the only alternative to rejecting it. Rather, there are at least three responses one can make to an apology: rejection, acceptance, or demurrer. This third option acknowledges that the apologizer has done nothing for which the need to apologize arises.[41] The second option is less indulgent, for it avows precisely what the third denies: that the actor *has* committed a wrong for which an apology is

38. While this is the dominant view, it is contested. For variants on the opposing view see, for example, MATTHEW H. KRAMER, WHERE LAW AND MORALITY MEET 249–94 (2008) (arguing for strict moral liability); John Gardner, *The Wrongdoing that Gets Results*, PHIL. PERSP., 2004, at 53, 69–86 (arguing for moral duties to succeed).

39. BERNARD WILLIAMS, *Ethical Consistency*, *in* PROBLEMS OF THE SELF 166, 175 (1973).

40. In a context such as this, "it wasn't your fault" is the idiomatic way to deny not merely that the actor is blameworthy, but also that he did anything wrong. I believe such an expression appropriate. As Matt Kramer rightly emphasized to me in private correspondence, his view accepts that *C*'s family should acknowledge *D*'s faultlessness or blamelessness, but not his purported lack of wrongdoing. Email from Matthew H. Kramer, Professor of Legal and Political Philosophy, University of Cambridge, to author (Apr. 15, 2010, 15:39 CST) (on file with author).

41. I am somewhat simplifying a yet more nuanced moral landscape. It might be that *D* does have a duty to apologize, as *P* of course would not, but that *C*'s family has a duty to affirm (in response to the apology) that *D* had no such duty. Morality and etiquette are tightly intertwined here, for while our fundamental obligation is to respect and express respect for other persons, we often have no better way

required. So insofar as "no harm, no foul" serves to accept an apology, its force or upshot is not to *deny* the commission of a foul, but to *affirm* it.

This understanding of the saying comports with its original usage. Longtime Lakers announcer Chick Hearn coined the expression in the 1960s to express the idea that referees should not call minor fouls that do not interfere with the flow of play.[42] The definition of the adage offered by Wiktionary is consistent with its use by both Hearn and my wife: "Encapsulation of the idea that although technically a breach of some code or law may have occurred there is no need for punishment . . . or retribution if no actual damage occurred."[43] It offers an example: "He parked in my space but as I was away at the time: no harm, no foul."[44] In contexts such as these, "no harm, no foul" is a slight misnomer. The underlying idea would be rendered more accurately, if less gracefully, as *no harm, no penalty, notwithstanding foul.* Call this reformulation NHNP.[45]

B. THE PUZZLE OF "NO HARM, NO PENALTY"

This lengthy explication of "no harm, no foul" is important because it makes more perspicuous a modest puzzle. Some norms and rules can be violated only by the causing of harm or injury, whereas other rules can be violated by proscribed conduct all alone, regardless of whether that conduct causes any further bad state of affairs. As a very rough generalization (albeit one subject to many exceptions), tort law (largely designed to ensure compensation for injury) requires harm for the commission of a foul, whereas criminal law (principally designed to deter commission of serious wrongs and to inflict retribution for blameworthy wrongdoing) does not.[46] Sports rules are similarly varied: some are defined such that their commission requires a proscribed result, but most

---

to satisfy that obligation than to conform to the complex social rituals designed to pattern respectful behavior.

42. *See, e.g.*, DAVID L. PORTER, BASKETBALL: A BIOGRAPHICAL DICTIONARY 203 (2005).

43. *No Harm, No Foul*, WIKTIONARY, http://en.wiktionary.org/wiki/no_harm,_no_foul (last modified Feb. 1, 2011). Omitted by the ellipses is the word "apology." I think that Wiktionary is wrong about that. In paradigmatic cases where no harm, no foul reasoning applies, there is a need for an apology, though the apology should be accepted. That is why it was my wife, not our babysitter, who uttered "no harm, no foul," and why she did so *after* he had apologized.

44. *Id.*

45. The difference between NHNF and NHNP can be captured by imagining that the actor repeats his infraction. Suppose, for example, that our babysitter left our house unlocked again, or your neighbor took your parking spot on another occasion. The aggrieved party might naturally invoke the earlier infraction when admonishing the wrongdoer for his behavior: "This is the second time you did such-and-such," we might say, reprovingly. Were "no harm, no foul" best understood literally, however, we would be wrong to raise that previous incident, for we would have already conceded that that initial event was no transgression at all. The first infraction counts from the vantage point of the second precisely because "no harm, no foul" means NHNP, not NHNF.

46. Kenneth W. Simons, *The Crime/Tort Distinction: Legal Doctrine and Normative Perspectives*, 17 WIDENER L.J. 719, 720, 727, 729 (2008).

aren't.[47]

"No harm, no foul" is often invoked to urge that a penalty not be imposed even when the relevant rules provide that harm is not required for the foul or for the consequent penalty, which is what Chick Hearn meant to express[48] and why Wiktionary has things right in emphasizing that, frequently, the adage is offered in response to "a breach of some code or law."[49] The puzzle, then, is this: given that rule makers know how to draft rules so that their violation does or does not require harm, and know how to specify that a harmless violation should incur no penalty, why would the nonrealization of harm be thought to warrant nonimposition of a penalty even in cases where the rule does not require harm?

The question is sufficiently important that I'll risk belaboring it. If in a *particular* case no penalty should be imposed because no harm was caused, notwithstanding that the conduct violated a rule that is not defined in terms of causing of harm, then why doesn't the reason for not imposing the penalty serve as well as a reason to rewrite the rule to require harm-causation in *all* cases? Put another way, is there an argument for not imposing a penalty on the commission of a foul in a particular case because, in *that* case, the foul caused no harm that allows for the possibility that the penalty should be imposed in *other* cases just

47. One sport that resolutely embraces the principle "foul, regardless of harm" is basketball: the victim of a shooting foul goes to the charity stripe even if his shot had gone in. *See* NAT'L BASKETBALL ASS'N, *supra* note 25. Golf is similar. For example, a player is disqualified for signing a scorecard that reports a lower score for any hole than actually taken. R&A RULES LTD. & THE U.S. GOLF ASS'N, RULES OF GOLF 57 (31st ed. 2007) (Rule 6-6(d)). Imagine the player who scores a three on hole twelve and a four on hole thirteen, but who accidentally switches the scores when recording them. The final recorded score is correct, but the player is nonetheless disqualified.

In football, similarly, most infractions incur penalties regardless of whether the infraction affected the play. For example, when a long run from scrimmage (or a long return of a punt or kickoff) is called back because of a hold or an illegal block by the offense (or the receiving team), it is not uncommon for the announcer to observe, in commiseration with the penalized team, that the infraction was particularly unfortunate because it occurred at a place on the field where the player held or blocked could not possibly have had an impact on the play. But the rule for pass interference is instructively different: "Contact that would normally be considered pass interference [is not pass interference if] the pass is clearly uncatchable by the involved players." *NFL Rules Digest: Pass Interference*, NFL.COM, http://www.nfl.com/rulebook/passinterference (last visited Apr. 28, 2011).

In soccer, the offsides rule is much like the uncatchable rule in football: a player who is in an "offside position" does not commit an "offence" if he is not "involved in active play" by interfering with the play or with an opponent, or by "gaining an advantage by being in that position." FÉDÉRATION INTERNATIONALE DE FOOTBALL ASS'N, LAWS OF THE GAME 2010/2011, at 31 (2010) (Law 11). But what about the player who simulates an injury with the intent of deceiving the referee? This is an offence for which a yellow card must be awarded regardless of whether the player has successfully deceived anyone. *Id.* at 115 (Law 12, Decision 5). In table soccer (foosball), "[s]pinning of the rods is illegal," INT'L TABLE SOCCER FED'N, RULE BOOK 11 (2007) (Rule 15), but "[s]pinning of a rod which does not advance and/or strike the ball does not constitute an illegal spin." *Id.* (Rule 15.2).

Most of the sports rules that make harm matter actually take the form NHNF, not NHNP. For present purposes, this difference is not material. The present point is that rule makers know how to specify that harm is a necessary condition either for a finding of violation or for the imposition of a penalty. Therefore, their failure to avail themselves of either option suggests that a given rule of conduct can be violated, and its violation penalized, even in the absence of a harmful consequence.

48. *See supra* note 42 and accompanying text.

49. *See supra* note 43 and accompanying text.

because of the foul and without regard for whether harm was caused? The plausibility of "no harm, no penalty, notwithstanding foul" depends on explaining how this no-harm-causing act-token of a proscribed act-type differs from other no-harm-causing act-tokens of that same proscribed act-type.

Briefly and quite generally, I think the answer (or part of it) is this: Were a penalty imposed solely for compensatory or restitutionary purposes, we would have no reason to enforce it when particular fouls end up being costless. Furthermore, we have affirmative reasons *not* to enforce it. Among other things, such penalties disrupt the flow of the contest and they handicap a competitor who has imposed no cost on his opponent. But, of course, penalties usually serve other functions too—most importantly, deterrence. This is why it is sensible to enforce penalties authorized for violations of act-types that are proscribed because of their tendency to cause harm, even in token situations where those act-types happen not to cause any harm.

Usually, but not necessarily always. On this account, nonenforcement might be warranted if, for context-specific reasons, enforcement of a penalty in a particular case would be unusually costly to the rule breaker (or to other interests), or if nonenforcement of the penalty on this occasion would weaken the deterrent force of the rule to an unusually small degree.

Let's recap. We started by asking what "no harm, no foul" means. We considered and rejected one oft-suggested answer: that an action is necessarily justified or not wrong if it fails to produce harm. Instead, I proposed that it is often used to accept an apology in circumstances when somebody did do something wrong but, happily, nothing bad came of it. We saw that this usage gives the lie to the expression because it means not that no foul has occurred but that no bad consequences should be imposed even though a foul *did* occur. And we saw that this formulation is curious. Because rules are sometimes crafted to make occurrence of harm a necessary condition for the foul and for any prescribed consequences, when a rule is crafted otherwise, the apparent implication is that any prescribed penalty should attach regardless of whether harm has occurred. Such a practice, despite a possible whiff of unfairness, can be easily defended on the ground that imposing a penalty regardless of harm improves deterrence of risky conduct in the future. Finally, we said that it might make sense to refrain from imposing a penalty even of such rules if, in a given case, no harm has occurred *and* either enforcement of a penalty in a particular case would be unusually costly to the rule breaker (or to other interests), or if nonenforcement of the penalty on this occasion would weaken the deterrent force of the rule to an unusually small degree.[50]

If this is right, then it should be apparent what more we need to justify

50. Of course, this analysis assumes all else remains equal. If, for example, enforcement of the prescribed penalty on this occasion would impose an unusually great cost on the rule breaker, but nonenforcement of the penalty would weaken the rule's deterrent force to an unusually great degree, then the conclusion suggested above might not follow. *See infra* note 76.

temporal variance in the enforcement of sports penalties. We need to establish that, at crunch time, penalizing harmless fouls would be unusual in one of the respects just mentioned. Before exploring whether that might be so, we should examine one small wrinkle. We have translated "no harm, no foul" (NHNF) into "no harm, no penalty, notwithstanding foul" (NHNP). But what if the conduct *does* cause harm? What if it causes some harm, but very little?

C. FROM "NO HARM, NO PENALTY" TO "LITTLE HARM, NO PENALTY"—OR NOT?

If prescribed penalties should sometimes not be enforced against the violator of a rule that is itself defined without regard to harm, when and because that particular violation caused *no* harm, does it also follow that there will be some occasions (albeit fewer) in which the penalty should not be enforced because the particular rule violation caused only *minor* harm? That is, if we are forced to choose one of two suboptimal alternatives—either (a) leaving a minor injury entirely uncompensated and thus allowing a rule breaker to enjoy the benefits of his violation, or (b) substantially overcompensating the victim and thus shifting a cost upon the rule breaker substantially in excess of the cost he would otherwise be allowed to impose on this victim—ought a sensible system ever to allow the injury to go unremedied?

One might think not. It is tempting to suppose that, as much as we might prefer perfect restitution, if the only options available to us are overpenalizing a rule breaker or undercompensating his victim, justice or fairness demands that we always select the former. We might say that, by violating the rule, an actor has "assumed the risk" that he'd be subject to a disproportionately excessive penalty, or that he forfeited his claim against a disproportionate penalty. Surely notions like guilt, fault, and innocence are relevant to our choice between suboptimal alternatives. The present question is whether they are always decisive.

The short answer is that they aren't—a proposition hard to conclusively establish, but easy enough to illustrate with well-settled legal practices that require victims of wrongdoing to simply lump it when the harm they have actually incurred is too slight.

One well-known example is contract law's material breach (or "substantial performance") doctrine, familiar to generations of law students from the leading case of *Jacob & Youngs, Inc. v. Kent*.[51] Jacob & Youngs contracted to build a Long Island home for a property owner at a total cost just north of $77,000. The owner specified that all pipe used in the building must be manufactured by the Reading Manufacturing Company. The contractor, probably inadvertently, installed pipe manufactured by other companies, though generally understood to be of quality and price equivalent to that made by Reading. After the house had been substantially completed, the owner discovered the discrepancy and refused

51. 129 N.E. 889 (N.Y. 1921).

to pay the remaining balance of $3,483. He instructed the contractor to remove the offending pipe and replace it with pipe that would conform to the contractual requirements—namely, Reading pipe. The New York Court of Appeals, in an opinion by Benjamin Cardozo, held for the contractor.[52]

There was no doubt that the contractor had breached his contract. The usual remedy for breach was to pay money damages sufficient to put the victim of the breach in the position for which he had contracted. But not in this case, said the court: because the breach was minor and full performance would have "meant the demolition at great expense of substantial parts of the completed structure,"[53] the court refused to recognize a forfeiture of the total sum remaining due to the contractor, a sum "grievously out of proportion" to the home owner's injury.[54] Instead, it held that the contractor only forfeited the difference in market value between the installed pipe and that specified by contract—which was to say, on the facts of this case, nothing.[55]

The "harmless error doctrine" governing appellate review reinforces the point—especially as applied to appeals from criminal defendants.[56] Suppose an appellate court is persuaded that a criminal defendant's legal rights were violated at the trial that resulted in his conviction. Say, for example, the trial judge improperly commented on the defendant's decision, protected by the Fifth Amendment privilege against self-incrimination, not to testify on his own behalf; or admitted out-of-court statements obtained from the defendant in violation of his right to counsel; or erroneously barred the defendant from representing himself. Possibly the most natural remedy would be to vacate the conviction and order a retrial. But trials are expensive and time-consuming. And because so many criminal trials are infected by minor errors, requiring retrial in all such cases would, in the aggregate, substantially increase the criminal courts' burdens, increasing delays across the board.[57] Furthermore, things may have changed since the initial trial that make a conviction unreasonably more difficult—memories might have become hazier, a key witness might have died. For all these reasons, courts and legislatures have concluded for nearly a century and a half—starting with the English Judicature Act of 1873[58]—that retrial should not be automatically ordered.

While the precise tests employed by the courts are complex and vary across jurisdictions, the common theme is simple. Except for a few constitutional violations that threaten core notions of procedural justice (prosecution in viola-

52. *Id.* at 892.
53. *Id.* at 890.
54. *Id.* at 891.
55. *Id.* at 891–92.
56. The rule and its history are summarized in Roger A. Fairfax Jr., *A Fair Trial, Not a Perfect One: The Early Twentieth-Century Campaign for the Harmless Error Rule*, 93 MARQ. L. REV. 433 (2009).
57. *See, e.g.*, *id.* at 436–37.
58. *See id.* at 435 & n.7 (citing Supreme Court of Judicature Act, 1873, 36 & 37 Vict., c. 66, § 48 (Eng.)).

tion of an individual's right to be free from double jeopardy, for instance, or the flat denial of a defendant's right to counsel), constitutional violations at trial will not be remedied by vacation of the conviction and an order of retrial if the appellate court concludes that the legal error did not likely contribute to the jury verdict—if, in the language of the doctrine, the error was probably "harmless." Although many aspects of the doctrine are controversial, the basic idea that a just system of criminal law need not remedy all violations of trial rights is rarely contested. But the doctrine is misnamed. The inquiry does not truly identify errors that are wholly harmless to defendants: if nothing else, the violation of a constitutional right is almost invariably injurious to the right holder's dignity and justified sense of entitlement. Rather, much like the superficially dissimilar material breach doctrine, it serves to withhold remedies that are "grievously out of proportion" to the injury sustained.[59] Put another way, the harmless error doctrine serves not to isolate errors that are truly harmless but to restrict the awarding of windfall remedies.

The lesson is simple. Just as *no harm*, *no foul* (NHNF) is frequently more accurately rendered as *no harm*, *no penalty* (NHNP), NHNP also includes *little harm*, *no penalty* (LHNP).

D. ON THE SAYING "IT COST US THE GAME"

In week six of the 2009 NFL season, the undefeated Minnesota Vikings hosted the 3–2 Baltimore Ravens in what turned out to be a thriller.[60] Up by 17 with ten minutes to go, the Vikings looked like they were on their way to a blowout. But second-year quarterback Joe Flacco led the Ravens to three quick touchdowns to take a 31–30 lead with under four minutes remaining. The ageless Brett Favre responded, driving the Vikings to a go-ahead field goal just inside the two-minute mark. At the end of regulation, Ravens kicker Steven Hauschka missed wide on a forty-four-yard field goal attempt that would have given the Ravens the victory. Not surprisingly, many fans bemoaned that the miss "cost us [the] game[]."[61] But Ravens running back Ray Rice, whose 194 yards from scrimmage were wasted in the loss, demurred: "We didn't lose that game because of Hauschka's miss . . . . If we start fast and put points on the board, our defense starts fast, I think the game is a totally different outcome."[62]

This is a sentiment commonly expressed across team sports. After some late-game mishap leads to a loss—a dropped pass or blown coverage, a pair of

59. *See supra* note 54 and accompanying text.

60. *See Baltimore Ravens v. Minnesota Vikings*, ESPN (Oct. 18, 2009), http://scores.espn.go.com/nfl/recap?gameId=291018016.

61. *E.g.*, It's NO GOOD!, Comment to *Steve Hauschka: 'It's Definitely One I Want Back*,' RAVENS INSIDER (Oct. 18, 2009, 9:03 PM), http://weblogs.baltimoresun.com/sports/ravens/blog/2009/10/post_4.html.

62. *Vikings Barely Stay Perfect, Hold off Flacco, Ravens*, CBSSPORTS.COM (Oct. 18, 2009) (internal quotation marks omitted), http://www.cbssports.com/nfl/gamecenter/recap/NFL_20091018_BAL@MIN.

missed free throws, a base-running error, or a referee's bad call—a large number of fans are sure to complain that the mistake "cost us the game." Other fans, players or coaches, will disagree. No, they'll protest, *that* play didn't cost us the game. Had we played better in other facets of the game—had we been able to convert on earlier trips into the red zone, or had we left fewer men on base, or had we capitalized on our power plays—we wouldn't have been in that situation in the first place.

The second assertion is surely right. Contrary to the adage, defeat has as many fathers as does victory. It can fairly be said about any number of plays that, had it gone differently, the loss would have been a win. But what about the first part of the protest? Does it follow from the fact that many players and coaches share responsibility for the loss that the particular poor play or bad decision did not cost the team the game? Was Ray Rice right that the Ravens "didn't lose that game because of Hauschka's miss"?[63]

Perhaps it depends on what we mean by "because of." Had Hauschka not missed, he would have made it.[64] And had he made it, the Ravens would have won, 34–33. So Hauschka's miss was a but-for cause of the Ravens' loss: but for the miss, they would not have lost; they would have won. In this fairly straightforward sense, the Ravens *did* lose because of Hauschka's miss. But—and this is very likely what Rice really meant—the Ravens didn't lose *only* because of Hauschka's miss. There were many but-for causes of their loss.

So if it is true that Hauschka's miss was a but-for cause of the loss, but wasn't the only such cause, why focus on it? I think that Rice and others who say similar things are urging that we really ought not to, that we focus on it only because it is more salient to us, perhaps because we know how things would have turned out had he done otherwise, whereas, although there are many other but-for causes, we forget about them, or don't know which they were. Rice is claiming that Hauschka didn't cause any more damage than did any other Raven who missed a play in that contest.

That is the question to investigate. More precisely, of course, we're interested in the more general question that this particular query exemplifies. The general question is whether the costliness of detrimental contest events varies depending on time of contest. And here's a way to focus the question. Suppose two missed forty-four-yard field goals—the first with time expiring in the first quarter, the second with time expiring in the fourth, both with the kicking team down 10–9. Are the two misses equally costly, or was one more costly than the other? And if one was the more costly, which?

The more common intuition, I venture, is that the latter miss is more costly than the earlier one. But many colleagues with whom I have spoken answer that the two are equally costly. I will argue that the latter is more costly than the

63. *Id.*

64. This assumes, of course, that the closest possible world to the actual is one in which the field goal is still attempted.

former. And, therefore, that there is a straightforward sense in which Hauschka's miss was especially costly precisely because of when it occurred. It's not just a matter of drama and atmospherics. Instead, generally speaking and all else equal, events have greater impact on the outcome of a game—good things contribute more to victory, and bad things are more costly—when they occur later in close contests. If that argument succeeds, we'll have the final piece of the explanation for slack in the calling of basketball fouls and similar infractions.

E. "CRUNCH TIME" AND THE VARYING MAGNITUDE OF OUTCOME-AFFECTING EVENTS

How ought we to think about the costliness of outcome-affecting events? I propose that the competitive costs or benefits of *any* game action, *X*, can be conceptualized in terms of the action's impact on a given competitor's probability of victory.[65] Assuming that the probability of victory ranges from 0 to 1, the

65. Philosophers and mathematicians of probability commonly distinguish two main types of probability, often characterized as subjective or epistemic on the one hand and objective on the other. The former "take probability to be concerned with the knowledge or belief of human beings," either actual or idealized in some respect. *See* DONALD GILLIES, PHILOSOPHICAL THEORIES OF PROBABILITY 2 (2000). "On this approach probability measures degree of knowledge, degree of rational belief, degree of belief, or something of this sort. . . . Objective interpretations of probability, by contrast, take probability to be a feature of the objective material world, which has nothing to do with human knowledge or belief." *Id.* The dominant view is that probability judgments regarding "singular events"—like whether this particular coin will land heads or tails on the next toss, or whether some particular individual will contract cancer within the next year—are invariably epistemic (because the "objective" probability of such events is either incoherent or always either 0 or 1). *See, e.g.*, D.H. MELLOR, PROBABILITY: A PHILOSOPHICAL INTRODUCTION 36 (2005). Because we are concerned here with the probabilities of singular events, our probability assessments must be epistemic. Given our familiarity and comfort with epistemic probability assessments, the fact that the probability estimates I'll be invoking are epistemic, not objective, will not, I believe, strike most readers as cause for concern. Nonetheless, I'd like simply to flag that scholarly debates over the "lost chance" doctrine in tort law, and particularly the law of medical malpractice, provide some grounds to be less sanguine.

Generally recognized in the United States and Canada but rejected in Britain, the "lost chance" doctrine arose as a response to the fact that negligent misdiagnosis or mistreatment will provide no grounds for recovery for a tort plaintiff who cannot establish by a preponderance of the evidence that the defendant's negligence caused him ordinary injury. *See, e.g.*, Steven R. Koch, *Whose Loss Is it Anyway? Effects of the "Lost-Chance" Doctrine on Civil Litigation and Medical Malpractice Insurance*, 88 N.C. L. REV. 595, 595–611 (2010). The doctrine provides, accordingly, that the negligent conduct (act or omission) that increases the probability of uncontroversial injury (usually death, bodily injury, or sickness) is *itself* a compensable injury. *Compare, e.g.*, Herskovits v. Grp. Health Coop., 664 P.2d 474, 479 (Wash. 1983) (finding a 14% reduction in plaintiff's chance of survival sufficient for recovery), *with* Hotson v. E. Berkshire Area Health Auth., [1987] 2 W.L.R. 287, 303 (appeal taken from Q.B.), *rev'd* [1987] A.C. 750 (H.L.) (finding a less than 50% decrease in plaintiff's odds of survival insufficient for recovery). Several scholars, Stephen Perry most influentially, criticize the doctrine essentially on the grounds that the increase in probability of injury that the doctrine assumes must be epistemic, not objective, and that a change in one's epistemic probability of injury cannot itself be an injury in the sense of a welfare setback. *See, e.g.*, Stephen R. Perry, *Protected Interests and Undertakings in the Law of Negligence*, 42 U. TORONTO L.J. 247 (1992). *Compare, e.g.*, Matthew D. Adler, *Risk, Death and Harm: The Normative Foundations of Risk Regulation*, 87 MINN. L. REV. 1293, 1346 (2003) (agreeing with Perry that risk damage is not a welfare setback), *with* Claire Finkelstein, *Is Risk a Harm?*, 151 U. PA. L. REV. 963, 995–97 (2003) (arguing otherwise). *See generally* Stephen Perry, *Risk, Harm, Interests, and Rights*, *in* RISK: PHILOSOPHICAL PERSPECTIVES

competitive cost or benefit of action X ranges from -1 to $+1$. Let us represent the impact of X on competitor A's probability of victory by $\varphi X(A)$. Now, let I_n-M stand for a token of infraction-type n committed by competitor M, and let P_n-M stand for a token of penalty-type n imposed on competitor M. Thus: $\varphi I_n-A(B)$ is the impact of a given infraction by A on B's probability of victory, and $\varphi P_n-A(A)$ is the impact on A's probability of victory of being assessed some given penalty. Lastly, because penalties for fouls in sports including basketball, football, and hockey are designed to deter, they must be (by and large) overcompensatory.[66] So: $|\varphi P_n-A(A)| > |\varphi I_n-A(B)|$.

(Tim Lewens ed., 2007); Stephen R. Perry, *Risk, Harm, and Responsibility*, *in* PHILOSOPHICAL FOUNDATIONS OF TORT LAW 321 (David G. Owen ed., 1997); Stephen Perry, *Harm, History, and Counterfactuals*, 40 SAN DIEGO L. REV. 1283 (2003).

A minimally adequate analysis of the problem of lost chances—what are better termed augmented probabilities of harm—is beyond the scope of this Article. Nonetheless, I hope that readers will appreciate that the problem of temporal variance in sport might both benefit from, and illuminate, doctrines of tort law that similarly depend upon changes in probabilities of the outcomes of which we are centrally concerned (outcomes like victories and deaths).

66. You might think that the common basketball practice of intentionally fouling when behind late in games—when stopping the clock is more important than not giving up points—is a counterexample to the claim in text: the competitive benefit from fouling (and the competitive cost of *being fouled*) is greater than the competitive cost of free throws to one's opponent. If it weren't, teams wouldn't keep doing it. In fact, though, the prevalence of strategic fouling in basketball and other sports (consider, for example, the defensive back in football who, knowing that he has been beaten, grabs a wide receiver when the ball is in the air, or even before it has been thrown) does not necessarily undermine my claim that penalties for prohibited conduct are designed to be overcompensatory. Rather, these practices may show either of two things.

First, it might be that intentional fouling late in games *is not prohibited.* If you drive southward from Marin County over the Golden Gate Bridge, you will be compelled to pay a toll on the far side. That toll is not a penalty, and your driving into San Francisco is not prohibited. The toll, while no doubt unwelcome to you, is a price exacted for permitted conduct, not a penalty or sanction imposed for prohibited conduct. On the distinction, see generally Robert Cooter, *Prices and Sanctions*, 84 COLUM. L. REV. 1523 (1984). In much the same way, the rules of organized basketball might be best understood as permitting teams to intentionally foul late in the game for a price. Although that might seem absurd at first blush, something can be said for it. While nobody much likes the stretching out of games with one team fouling the other on each possession, to prohibit the practice might be tantamount to affirming that it's preferable to let a team with a lead pass and dribble out the clock. It's not crazy to think that would be worse.

But maybe you think that would not be worse. Maybe it would be better to prohibit fouling late in games even if it means that one team gives up any realistic chance to win. If fouls are always prohibited and never meant to be permitted-for-a-price, then we come upon the second possible lesson: the rules of professional basketball are not, in this respect, well designed. In order to cut down on strategic end-of-game fouling, the NCAA increased the penalty for intentional fouls in 1987 to provide that the victimized team gets two free throws and possession out of bounds. *See History of NCAA Basketball Rule Changes*, ORANGEHOOPS, http://www.orangehoops.org/NCAA/NCAA%20Rule%20Changes.htm (last visited Apr. 26, 2011). Maybe the NBA should follow the NCAA's lead. Then again, the seeming fact that college referees rarely call late-game strategic fouls as "intentional," and therefore refrain from awarding the victimized team post-free-throw possession as the formal rules would require, might suggest that the logic of the game creates substantial pressure to treat late-game strategic fouls as permitted-for-a-price.

There is a large philosophy-of-sports literature on the ethics of strategic fouling, much of it cited and discussed by Robert L. Simon, *The Ethics of Strategic Fouling: A Reply to Fraleigh*, 32 J. PHIL. SPORT 87, 88 (2005) (arguing that strategic fouling is not always unethical), and Warren P. Fraleigh,

It should be apparent on little reflection that both $\varphi I_n - A(B)$ and $\varphi P_n - A(A)$ are context variant to at least some extent. For example, they vary depending on the score differential at the time of the event: each is greater when the score is tied than when either team enjoys an insurmountable lead. The claim I want to add is that, holding all else constant, they are also temporally variant. This is for the simple reason that, holding closeness of contest constant, scores and missed scoring opportunities have greater impact on the outcome of a close game the closer they occur toward the game's end.[67]

Here's an illustration: Suppose that the Lakers are called for a shooting foul on a missed basket by the Celtics in the first twenty seconds of a scoreless scheduled forty-eight-minute game. Suppose that the shot missed the basket and backboard and sailed out of bounds. Absent the enforcement of a penalty, the Lakers would have been awarded the basketball, and the score would have remained 0–0. Assuming that the teams were evenly matched and playing at a neutral site, the Lakers' probability of victory absent a penalty would have been, let us suppose, 0.5. The awarding of two free throws to the Celtics as a penalty for the Lakers' foul has an expected value of 1.5 points.[68] The Lakers' probability of victory when down 1.5 with 47:40 remaining is, let us imagine, 0.495. The cost of the penalty to the Lakers is -0.05, a small cost indeed. Now assume the same facts, except that the foul is called and penalty assessed with time expiring and the Lakers ahead by one point. The Lakers' probability of victory absent the penalty would be exactly 1.0: the game would have ended at that moment with the Lakers enjoying a razor-thin lead. The probability, post-penalty, that the Lakers will win is the probability that the Celtics will miss both free throws plus the probability that the Celtics will hit only one of the two free throws and then lose in overtime, which is 0.25.[69] The cost of the penalty to the Lakers under these circumstances is a whopping -0.75.

If that is so, then we can see the reason for preferring that the penalty not be imposed for this particular infraction: *We want the outcome of athletic contests*

Intentional Rules Violations—One More Time, 30 J. PHIL. SPORT 166 (2003) (arguing that it is). I favor Simon's analysis, though I add this quibble: because Simon explicitly invokes the price–sanction distinction, he ought not to agree that "[s]trategic fouls occur when a competitor in an athletic contest deliberately and openly breaks a rule expecting to be penalized and with willingness to accept the penalty, in order to obtain a strategic advantage in the contest." Simon, *supra*, at 87. If there are times during the course of a contest that strategic fouling is best conceptualized as permitted-for-a-price, then the strategic fouler does not, on those occasions, "break a rule."

67. The closeness of a contest is a function, *inter alia*, of the score differential and the distance remaining to contest completion. *See supra* note 32. For a constant score differential of n (greater than zero), the contest is closer at $t1$ than at $t10$. What I mean to keep constant is closeness of contest, not score differential.

68. *See* John Branch, *For Free Throws, 50 Years of Practice Is No Help*, N.Y. TIMES, Mar. 3, 2009, at A1 (observing that the average free-throw shooting percentage has remained steady at approximately 75% for more than fifty years).

69. Assuming a 0.75 probability of converting each free throw, the probability of missing both is 0.25×0.25, and the probability of converting exactly one is $0.75 \times 0.25 \times 2$. Assuming that each team has an equal probability of winning if the game reaches overtime, the Lakers' probability of winning if the Celtics are awarded two free throws is $(0.25 \times 0.25) + (0.75 \times 0.25 \times 2 \times 0.5) = 0.25$.

to depend (insofar as possible) upon the competitors' relative excellence in executing the particular athletic virtues that the sport is centrally designed to showcase and reward,[70] and a sanction of this magnitude would make the outcome too dependent on the less important (though not unimportant) ability to refrain from any bodily contact. Let me be very clear: what I have just said is not to claim that this latter excellence is something that, in the nature of things, no sport could wish most to valorize; I am claiming only that, in the sport of basketball as we know it, this particular excellence does not rank so highly among the excellences that we wish to feature and encourage.

Or consider another illustration, moving from basketball to football. The Cowboys are beating the Giants 21–20 with one minute remaining in the fourth quarter and the Giants facing fourth-and-ten from their own forty-yard line. Giants quarterback Eli Manning throws the ball forty yards downfield, where a Cowboys' cornerback modestly interferes with the Giants' receiver. Even absent any interference, the catch would have been very difficult—say, a 0.3 probability. The modest physical contact made a reception more difficult but still possible—say, a 0.2 probability. Either way, had the receiver caught the ball, his momentum would have carried him out of bounds at the Cowboys' ten-yard line. In the event, the pass falls incomplete. On these assumptions—stylized but not fanciful—the interference imposed a cost of -0.1 likelihood of reception. A reception at this field position and time, let us say, would have given the Giants a 0.96 probability of victory. An incompletion gives possession to the Cowboys and thus leaves the Giants with a mere 0.01 probability of victory.

Should the foul be called and the penalty—awarding the Giants possession and a first down at the spot of the infraction[71]—enforced? Some informed observers would say no, arguing that during crunch time referees should give players a little more latitude for physical contact.[72] Without fully resolving whether, on balance, such arguments are correct, the analysis to this point at least makes sense of why they might be. The contest-outcome cost of this defensive pass interference to the Giants ($\varphi \mathrm{I}_n - C(G)$) is -0.095. The contest-outcome cost of the penalty to the Cowboys ($\varphi \mathrm{P}_n - C(C)$) is -0.95. When a penalty would have such a substantial expected effect on the game's outcome and when the unpenalized infraction would not have an effect of roughly similar magnitude, we might think that the game goes better—is fairer and more satisfying—if the play on the field stands.[73]

---

70. For the now-classic discussion in the philosophy-of-sport literature, see J.S. Russell, *Are Rules All an Umpire Has To Work With?*, 26 J. PHIL. SPORT 27, 35 (1999) (urging that "[r]ules should be interpreted in such a manner that the excellences embodied in achieving the lusory goal of the game are not undermined but maintained and fostered").

71. *NFL Digest of Rules: Pass Interference*, NFL, http://www.nfl.com/rulebook/passinterference (last visited Jan. 23, 2011).

72. *See supra* note 29.

73. An alternative to ignoring the infraction entirely, of course, would be to impose a more modest penalty. Possibly, the NFL makes things unnecessarily hard on itself by insisting that defensive pass interference always be penalized at the spot of infraction, instead of allowing for more modest yardage

In short, the expected outcome-affecting magnitude of an outcome-affecting event is greater toward the end of a contest, holding closeness of contest constant, because there are fewer opportunities to counter the impact of that event.[74] Another way to think about the point is this: the less time remaining in

---

mark-offs for minor interference. The NCAA penalizes defensive pass interference at the spot of infraction only if it occurs within fifteen yards from the line of scrimmage; if the infraction occurs farther downfield, the defense is penalized fifteen yards. *See* NAT'L COLLEGIATE ATHLETIC ASS'N, FOOTBALL: 2009–10 RULES AND INTERPRETATIONS 105 (Ty Halpin ed., 2009) (Rule 7-3). But it's unlikely that the NFL will find this proposal congenial, for in the long-running battle between "lumping" and "splitting" (categorizing heterogeneous phenomena into few broad classes that emphasize similarities or into many narrow classes that emphasize differences), *see* Bradley C. Karkkainien, Reply, *"New Governance" in Legal Thought and in the World: Some Splitting as Antidote to Overzealous Lumping*, 89 MINN. L. REV. 471, 479 (2004), the league's recent rescission of the five-yard face-masking penalty in favor of making all such infractions punishable by fifteen yards suggests that it has cast its lot with the lumpers, *see* John Clayton, *Owners Table Reseeding Playoffs Proposal; Pass Other Rules*, ESPN (Apr. 2, 2008, 7:08 PM), http://sports.espn.go.com/nfl/news/story?id=3325273. And this isn't obviously crazy. Sensible system designers will frequently make available a more limited range of penalty options than might initially strike us as possible and desirable—perhaps to mitigate risks of over- or under-deterrence, or to reduce the time and expense of rule application, or to serve other reasonable systemic goals. For example, while most U.S. jurisdictions authorize sentencing officials to impose any sentence for voluntary manslaughter between two broadly spaced poles (say, between two and twenty years), California authorizes only three possible sentences: three, six, or eleven years. CAL. PENAL CODE § 193(a) (West 2010). Closer to home, law students might contrast the thirty-two-interval grading system used at the University of Chicago (any numerical score from 155–186, inclusive) with the four-interval system employed at Yale (Honors/Pass/Low Pass/Fail). *Compare Key to Transcripts of Academic Records*, U. CHI., http://registrar.uchicago.edu/policies/transcript_key.shtml (last visited Jan. 25, 2011), *with Grades*, YALE L. SCH., http://www.law.yale.edu/academics/jdgrades.htm (last visited Jan. 25, 2011).

74. Of course, things are not always so simple, and some dynamics might cut in just the opposite direction. Most notably, insofar as early scores might have especially pronounced effects on game strategy and psychology, one might predict that, holding closeness of contest constant, points scored earlier in contests are likely to have a greater effect on outcome than points scored later. The effect of early scores on game progression is ultimately an empirical question and likely to vary substantially across different sports. Nonetheless, I suspect that the effects of early scores are less regular than this surmise suggests. Sometimes the team that goes down early will become demoralized, other times it will become refocused; sometimes the team that goes up early will play with greater confidence, other times it will become complacent or sloppy. But if you're losing, you simply can't prevail unless you score more than your opponent in the time remaining. And the less time that remains, the harder that it is to do.

A second qualification to the claim is best presented with an illustration. In a high-scoring matchup between two football teams with explosive offenses and weak defenses, it may often appear that whichever team gets the ball last will win. At some point in that game, a go-ahead score by Team *A* might maximize its probability of victory if the score comes with more, rather than less, time remaining. Team *A*'s best bet, of course, is to leave Team *B* with insufficient time to score. But failing that, Team *A* is better off leaving its opponent with too much time to exhaust, thus increasing the likelihood that if Team *B* does score, Team *A* will be left with the final possession. I suspect that the general phenomenon that this example instantiates is fairly unusual but cannot examine its contours or incidence here. In the meantime, we might dub the phenomenon "The Upper Hand Caveat," because it depends upon the nonlinear structure of alternating opportunities found in the familiar method for determining which of two captains picks first in selecting teams for sandlot baseball. As one authority explains the ritual:

> One puts a hand around the bat near the fat end, then the other puts a hand around the bat just above his hand. This goes on, hand over hand, until the bottom of the bat is reached and there is no room for another hand. The last hand on the bat wins the contest (although the loser does

the contest, the greater the effect of each unit change in score on "closeness of contest," and thus the greater the effect expected on outcome of each unit change in probability of a score.[75] "Crunch time" is just that period when everything matters more.

It is true that $\varphi I_n - A(B)$ and $\varphi P_n - A(A)$ both increase in crunch time and that the ratio between the two probably remains constant. However, the absolute magnitude of the difference between the two increases. Therefore, because $\varphi P_n - A(A)$ may substantially exceed $\varphi I_n - A(B)$, the change in the outcome-affecting difference between allowing the infraction to go unrectified and imposing a penalty might well justify heeding the call to "let 'em play"—certainly in cases of NHNP, and very possibly in cases of LHNP, too.[76]

III. TWO KINDS OF FAULTS, MANY KINDS OF RULES

At first blush, we might suppose that the analysis of Part II applies, *mutatis mutandis*, to foot faults in tennis and therefore that McEnroe's position is vindicated:[77] tennis officials should call foot faults less strictly at crunch time. And if that's true, it's a small step (sorry) to the conclusion that the penalty for foot faults should not have been imposed on Serena Williams, even assuming her foot did touch or slightly cross the baseline. But the conclusion that the analysis of Part II applies to foot faults would be premature. It could be that foot

---

have the chance to delicately grasp with his fingertips whatever little wood is left and twist it around his head, winning if he can hold on to the bat while doing this three times).

ROBERT HENDRICKSON, THE FACTS ON FILE: ENCYCLOPEDIA OF WORD AND PHRASE ORIGINS 280 (1997).

75. Peter King recognized that events gain in significance in inverse proportion to time remaining when analyzing Bill Belichick's much-discussed decision to go for it on fourth-and-two from his own twenty-eight yard line with 2:08 remaining, up 34–28 against Indianapolis in week ten of the 2009 season. *See* Peter King, *No Matter Which Way You Dissect It, Bill Belichick Made the Wrong Call*, SPORTS ILLUSTRATED (Nov. 16, 2009, 5:57 AM), http://sportsillustrated.cnn.com/2009/writers/peter_king/11/15/mmqb/index.html. The Patriots didn't get the first down; the Colts took over on downs and scored the winning touchdown with thirteen seconds remaining. King is critical of the decision, urging that New England should have punted instead. *Id.* His criticism was misguided. For a powerful defense of Belichick's decision, see Frank Frigo, *The Anatomy of a (Fourth-Down) Decision*, FIFTH DOWN (Nov. 25, 2009, 7:00 PM), http://fifthdown.blogs.nytimes.com/2009/11/25/the-anatomy-of-a-fourth-down-decision/. But in the course of his analysis, King makes an astute observation. Acknowledging that Belichick successfully made a similar call earlier in the season, King notes a critical difference:

> Against Atlanta in Week 3, there was a play something like this. New England had fourth-and-one at its 24 late in the third quarter, up 16–10. Sammy Morris ran for two yards, first down, and the Patriots went on to kick a field goal on the drive. But that was one yard, not two, and even if it had failed and the Falcons got the ball and scored, *the Patriots would have had an entire quarter to rectify things*.

King, *supra* (emphasis added).

76. I have focused on the respect in which the overcompensatory effect of a penalty can change from one moment in the contest to another. I have not addressed changes in the deterrent value of a penalty, though I acknowledge that such changes might be at least partially confounding. A full analysis of the dynamic interaction between these considerations is beyond the scope of this Article but would have to be undertaken before we could reach all-things-considered conclusions with any confidence. Recall my earlier declaration of this Article's more modest ambitions.

77. *See supra* note 5 and accompanying text.

faults in tennis differ from fouls and similar infractions in basketball, football, and comparable sports[78] in ways that make a difference. This Part advances two arguments: first (in sections III.A through III.C), that foot faults *do* differ in a way that matters; and second (in sections III.D through III.F), that temporal variance in their enforcement can nonetheless be defended on alternate grounds.

A. FROM THE HARDWOOD TO THE DIAMOND

I claimed in the Introduction that the case for temporal variance is made difficult by clear cases for invariance. My example was strikes in baseball. Nobody, I supposed, would think that umpires should cut pitchers a little more slack toward the end of a close game. But was I too quick? The famed Harvard-zoologist-cum-Yankees-fan Stephen Jay Gould would have us believe so. Here's his account of Don Larsen's perfect game in the 1956 World Series—the only perfect game or no-hitter in postseason play until Roy Halladay's no-hitter in the 2010 playoffs[79]—penned on the death of that game's home plate umpire, Babe Pinelli:

---

78. Hockey fans will notice that, although I have mentioned the sport a small handful of times, I haven't weighed in on the much-debated questions of whether power plays should be awarded less liberally in overtime than during regulation or in the playoffs than during the regular season. There's a good reason for that: I don't watch a lot of hockey. So I'm offering my four cents in a footnote in the hope that, if what I say here is unusually daft, fewer readers will notice.

First, it seems to me that the analysis provided so far does offer support for the view that referees should swallow their whistles in overtime. Second, the analysis probably also lends credence to the thought that the game should be called less tightly during the playoffs than during the regular season. We have thus far treated the relevant unit as the individual game: we want the outcomes of individual games to reflect the competitors' relative success in realizing the sport's core athletic excellences and in overcoming the sport's central challenges. But we could as well consider the unit of competition to be the season. Insofar as teams are competing for season success—in the NHL, winning the Stanley Cup—then whatever argument might support temporal variance for overtime relative to regulation would translate fairly straightforwardly to support temporal variance for the playoffs relative to the regular season.

Third, and on the other hand, power plays in hockey are arguably different from all of the penalties discussed so far. Free throws in basketball and yardage mark-offs in football might be necessary evils. There is nothing particularly exciting or valuable about either; the leagues would do away with them both but for the need to compensate victims of infractions for the competitive harm that has been done them and to deter future such infractions. At least some people, however, view power plays differently. Indeed, some have speculated that when, in 2006, NHL Commissioner Gary Bettman warned referees not to call the playoffs more loosely, he was motivated in part by the view that power plays are positive goods, not just necessary evils, because they increase scoring and scoring is always exciting. *See, e.g.*, Ritch, *OK—So Bettman's "Cracking Down,"* AM. HOCKEY FAN (Apr. 21, 2006, 12:38 PM), http://americanhockeyfan.blogspot.com/2006/04/ok-so-bettmans-cracking-down.html.

Fourth and finally, it appears to me that much of the debate about calling penalties in the playoffs is not principally about temporal variance. I think—and am open to being corrected—that many of the folks who urge that the refs should "let 'em play" in the playoffs would prefer that they "let 'em play" during the regular season, too. These fans simply prefer a more physical style of hockey. They focus their appeals on the playoffs not principally because they think the playoffs are meaningfully different, but in conformity (conscious or not) with the maxim that one ought to pick one's battles.

79. Dave Sheinin, *Roy Halladay No-Hitter Lifts Phillies Past Reds in Game 1*, WASH. POST (Oct. 7, 2010, 12:01 AM), http://www.washingtonpost.com/wp-dyn/content/article/2010/10/06/AR2010100606427.html.

> Babe Pinelli was the umpire in baseball's unique episode of perfection—a perfect game in the World Series. It was also his last official game as arbiter—Oct. 8, 1956. Twenty-seven Dodgers up; 27 Bums down. The catalyst was a competent, but otherwise undistinguished Yankee pitcher, Don Larsen.
>
> The dramatic end was all Pinelli's, and controversial ever since. Dale Mitchell, pinch hitting for Sal Maglie, was the 27th batter. With a count of one ball and two strikes, Larsen delivered a pitch low and outside—close, but surely not, by any technical definition, a strike. Mitchell let the pitch go by, but Pinelli didn't hesitate. Up went the right arm for called strike three. Out went Yogi Berra from behind the plate, nearly tackling Larsen in a frontal jump of joy.
>
> "Outside by a foot," groused Mitchell later. He exaggerated, for it was outside by only a few inches, but he was right. Babe Pinelli, however, was more right. A man may not take a close pitch with so much on the line. Context matters. Truth is a circumstance, not a spot.
>
> . . . .
>
> Truth is inflexible. Truth is inviolable. By long and recognized custom, by any concept of justice, Dale Mitchell had to swing at anything close. It was a strike—a strike low and outside. Babe Pinelli, umpiring his last game, ended with his finest, his most perceptive, his most truthful moment. Babe Pinelli, arbiter of history, walked into the locker room and cried.[80]

Is Gould right that Pinelli should have called that nonstrike a strike? Is this a case for temporal variance? I cannot do justice to Gould's remarkable essay in this space, so it must suffice to make two brief points. First, whereas the case for temporal variance I have sketched depends entirely on an assessment of the costs and benefits to the competitors, Gould rightly draws attention, in addition, to the effect that calls and no-calls can have on others—fans at the time and even, we might say, posterity. Second, and nonetheless, if my informal polling of students and colleagues provides reliable guidance, Gould is in a very small minority. Most of those I have asked think that strikes and balls are not appropriate candidates for temporal variance even here, in what we might reasonably suppose is the best-case scenario for it. If Mitchell "had to swing" at a close 1–2 pitch—if a batter "may not" take a close pitch in that setting—that's because it's unwise to risk a mistake about the pitch's location, either by himself or by the umpire. The command is prudential, not normative. Accordingly, that Mitchell should have been aware that the umpire might *erroneously* call a near miss a strike doesn't mean that the umpire should have done so purposefully.

---

80. Stephen Jay Gould, Op-Ed., *The Strike That Was Low and Outside*, N.Y. TIMES, Nov. 10, 1984, at 23. Gould's piece was mistitled. Video of the play makes clear, as the body of the essay suggests, that if the pitch wasn't the right height, it was too high, not too low—as Gould subsequently recognized. *See* STEPHEN JAY GOULD, TRIUMPH AND TRAGEDY IN MUDVILLE: A LIFELONG PASSION FOR BASEBALL 47 (2003); TheMLBhistory, *Don Larsen's Perfect Game Biography*, YOUTUBE (Dec. 22, 2010), http://www.youtube.com/watch?v=crTFMCkG7cU&t=02m50s.

Or, to put the thought in a nutshell: had Dale Mitchell been Ted Williams, custom and justice would not have demanded that he swing at a pitch he knew to be outside.[81]

If this is so, two tasks remain. First, we must examine whether the analysis developed in Part II that lends support for temporal variance in the enforcement of infractions like fouls in basketball can explain why slack is not appropriate in the calling of balls and strikes. If it can't, then we have reason to worry that we have someplace gone astray. But if it can, we are poised to undertake the second and final task: to determine whether the analysis already advanced, refined or supplemented as may prove necessary, applies to foot faults. Speaking loosely, the task will be to determine whether foot faults in tennis are more like fouls in basketball or like balls in baseball.

B. TWO KINDS OF RULES

The rationale for temporal variance in the enforcement of penalties for fouls relied upon "no harm, no penalty" reasoning.[82] We said that there are times when it might better serve the objectives of competitive sports to refrain from enforcing a penalty despite the occurrence of an infraction. That's because the competitive costs of an infraction and of the sanction or penalty that it begets are temporally variant and the latter can become, at game's end, very much greater than the former. Yet assessing the competitive costs of these two things—the infraction and the sanction—seems impossible in the case of balls and strikes. It's impossible because the denomination of a pitch as a "ball" is not properly conceptualized as the penalty for an infraction; the concepts of infraction and penalty just don't apply here.

Hart, in critiquing the Austinian command theory of law, argued that some consequences of rule violations are not actually sanctions.[83] Most of the rules of the criminal law impose duties and threaten sanctions for their violation. But other legal rules, like those specifying the conditions for valid wills or contracts, are of a different sort. These, Hart proposed, are "power-conferring rules"—rules that (somewhat simplified) provide that "[i]f you wish to do this, this is the

---

81. The Splendid Splinter's extraordinary vision and command of the strike zone—and the respect they earned him from the men in blue—is often illustrated with the story of the young catcher who complained that a close pitch was miscalled a ball. "Son," the ump is said to have replied, "when the pitch is a strike, Mr. Williams will let you know." ZACK HAMPLE, WATCHING BASEBALL SMARTER 122 (2007) (internal quotation marks omitted). In fact, many authorities attribute the quote to Hall of Fame umpire Bill Klem—and about Rogers Hornsby, not Ted Williams. *See, e.g.*, *Rogers Hornsby Quotes*, BASEBALL ALMANAC, http://www.baseball-almanac.com/quotes/quohorn.shtml (last visited Jan. 19, 2011). But that's nitpicking: Williams had a pretty fair eye, and the umps knew it.

82. *See supra* sections II.A–C.

83. Austin conceived of laws as sanction-backed commands of a person or determinate body of persons—the "sovereign," who receives habitual obedience from the bulk of the populace and who does not habitually obey anyone else. *See* JOHN AUSTIN, THE PROVINCE OF JURISPRUDENCE DETERMINED x–xi (Hackett Publ'g Co. 1998) (1832). In addition to the criticism developed in the text, Hart objected, *inter alia*, that Austin's analysis could not make sense of the continuity of a legal system or of the normative character of law. Hart, *supra* note 13, at 18–76.

way to do it."[84] In the case of rules that impose a duty, he explained,

> we can distinguish clearly the rule prohibiting certain behaviour from the provision for penalties to be exacted if the rule is broken, and suppose the first to exist without the latter. We can, in a sense, subtract the sanction and still leave an intelligible standard of behaviour which it was designed to maintain.[85]

But the distinction between the rule and the sanction is not intelligible in the case of power-conferring rules. It makes sense to say "do not kill" even when we leave off the part about what happens if you do. In contrast, we know we're leaving something critical out of the picture if we say "get two witnesses" but don't explain that the will is invalid otherwise.

The Hartian analysis of power-conferring rules helps to explain why balls and strikes in baseball feel very different from the infractions we have considered in basketball and football. In the case of the latter, we can sensibly ask whether some type of contact ought to be proscribed (thus denominated as a "foul") and, if so, whether the penalty attached to commission of the foul—two free throws, say, or ten yards—is too great (or too small). But every pitch is either a ball or a strike. The logical consequence of a pitch being outside the strike zone is that it's a ball. We can sensibly ask whether the strike zone is too small (or too large), or whether the number of balls that constitutes a walk is too great (or too small),[86] or whether *any* number of balls should result in the award of a base;[87] however, it seems nonsense to ask whether a pitch being a ball is too high a price for it narrowly missing the strike zone: that the pitch is a ball is just what it *means* for it not to be a strike.

In short, we might provisionally endorse the following conclusion: Contra Gould, balls and strikes are not proper candidates for temporal variance because (1) temporal variance depends upon the widening of a gap between the competitive cost of an infraction and the competitive cost of the penalty it incurs, but (2) there is no such gap between nonconformity with a power-conferring rule and the consequences that attach, and (3) the rules governing balls and strikes are power-conferring rules (or something of a sufficiently close type).[88]

---

84. Hart, *supra* note 13, at 28 (internal quotation marks omitted).

85. *Id.* at 34–35.

86. In fact, in the early days of baseball, more balls were required to make out a walk. Five successive rule changes adopted between 1879 and 1889 dropped the number from nine to the current four. DAVID NEMEC, THE OFFICIAL RULES OF BASEBALL ILLUSTRATED 22 (2006); GLEN WAGGONER ET AL., SPITTERS, BEANBALLS, AND THE INCREDIBLE SHRINKING STRIKE ZONE: THE STORIES BEHIND THE RULES OF BASEBALL 114 (rev. ed. 2000) ("In 1879 it took nine balls . . . . By 1889 it was down to four . . . .").

87. *See* NEMEC, *supra* note 86 (noting that early baseball had balls before it had bases on balls).

88. The parenthetical is intended to signal that I am not wholly committed to this particular typology of rules. Perhaps, for example, the roughly analogous distinction between regulative and constitutive rules, most prominently associated with John Searle and featured predominantly in the philosophy of sports literature, might provide the more useful analytical framework. I am open to other possibilities as well. Roughly, "regulative rules regulate antecedently or independently existing forms of behavior"

C. TWO KINDS OF FAULTS

If the analysis of balls and strikes is correct, then the question whether temporal variance would have been appropriate in the Williams–Clijsters match might depend on how we conceptualize the rule governing foot faults: as duty-imposing or power-conferring (subject to the qualification in the preceding footnote). If the rules command that the server not step on or past the baseline, on pain of the serve being declared null, then the reasoning that supports temporal variance in the calling of fouls would seem to be available, providing a possible basis for temporal variance in the calling of foot faults too. If, instead, the rules confer upon the server two opportunities (lets aside) to put the ball in play, and specify that the only valid way to put the ball in play is for the server not to step on or past the baseline, then the analysis proffered in Part II would not apply and we would have, thus far, no basis for temporal variance.

Which is the better conceptualization of the relevant tennis rules is, I think, more open to argument than one might suppose. There is no simple test that straightforwardly or uncontroversially resolves the question.[89] So let's start not with foot faults but with the more common type of service fault—the fault produced by the failure to strike the served ball into the diagonally opposite service court. Call this a "zone fault."

It seems as clear as these things can be that the rules of tennis are not rightly understood to impose upon a server the duty to strike the ball into the service court. To be sure, servers are under a duty to *try* to put the ball lawfully or validly into play, for playing a competitive game involves assuming an obliga-

---

whereas constitutive rules "do not merely regulate, they create or define new forms of behavior." JOHN R. SEARLE, SPEECH ACTS 33 (1970). And they create new forms of behavior—what Searle sometimes calls "institutional facts"—by assuming the form "*X* counts as *Y* in context *C*." *Id.* at 51–52 (internal quotation marks omitted). Thus, for example, moving the king two squares toward a rook and moving that rook to the square over which the king has crossed counts as castling in chess. *See generally* JOHN R. SEARLE, THE CONSTRUCTION OF SOCIAL REALITY (1995); RAIMO TUOMELA, THE PHILOSOPHY OF SOCIAL PRACTICES: A COLLECTIVE ACCEPTANCE VIEW (2002). For criticisms of the regulative/constitutive distinction, see, for example, Christopher Cherry, *Regulative Rules and Constitutive Rules*, 23 PHIL. Q. 301, 309 (1973); Frank Hindriks, *Constitutive Rules, Language, and Ontology*, 71 ERKENNTNIS 253, 253 (2009).

89. The text of the relevant rules is always a good place to start. Rule 18, recall, is written in deontic terms, providing, *inter alia*, that "[d]uring the service motion, the server shall not . . . [t]ouch the baseline or the court with either foot." INT'L TENNIS FED'N, *supra* note 6. Furthermore, Rule 19.a—by providing that "[t]he service is a fault if . . . [t]he server breaks rules 16, 17 or 18"—reinforces the idea that the server has a duty not to step on the line, for, strictly speaking, one does not "break" a power-conferring rule by failing to comply with its dictates. *Id.* However, we should not place too much weight on constructions that are unlikely to have received much conscious thought by the drafters. After all, Rule 17—a second rule that Rule 19 contemplates being broken—employs similar deontic language in providing that "[t]he service shall pass over the net and hit the service court diagonally opposite." *Id.* Yet no tennis player would think himself under a duty to hit the diagonally opposite service court. To the contrary, all players and fans understand that the server is empowered to put the ball into play by serving into the correct space. Because the plain function of Rule 17 is to confer a power, not to impose a duty, we must characterize it as a power-conferring rule, the somewhat infelicitous language notwithstanding. We ought not then invest much faith in the ordinary meaning of the text of Rule 18 either.

tion to compete. If your opponent doesn't even try to do the things necessary to score points, you will rightly object, notwithstanding that her failure to attempt to satisfy the power-conferring rules all but assures you of victory. However, a sincere but unsuccessful effort to strike the ball into the specified zone is not a violation of an obligation. It is an invalid serve—a "fault"—for the same reason that a will signed by only one witness is legally invalid and that an unswung-on pitch that misses the strike zone is a ball: the actor has failed to do what is specified to perfect a power. If, contrary to Gould, Babe Pinelli should not have punched Dale Mitchell out on Larsen's 1–2 pitch,[90] then a line judge should call any serve that's long or wide a service fault regardless of game context.

Although the issue is more debatable, on balance it seems sensible to characterize the rules defining foot faults as power-conferring as well. In order to successfully or "validly" put the ball into play, thus giving oneself an opportunity to win the point, the server must do several things: (1) start behind the baseline, (2) strike the ball before stepping on or over the baseline, and (3) by striking the ball, cause it to land in the service court diagonally opposite. We might say these are three components of the rule that defines a valid serve. A failure to complete any of these three actions is just a failure to perfect the power conferred upon the server; none is a violation or an infraction.

That seems right.[91] But here's the puzzling thing. If foot faults are also governed by power-conferring rules and if temporal variance could be defended only on the analysis developed to this point, then we should expect foot faults to be immune from temporal variance, just as are zone faults. But widespread intuitions are more equivocal. I have not run across anybody who is tempted to apply temporal variance for zone faults. If, facing match point, the server hits a second service wide by a smidgen, well them's the breaks and that's the match. And yet we have already seen that some—John McEnroe, for example[92]—believe that foot faults should be enforced with temporal variance. Just as revealingly, many more feel that the temporal variance of foot faults is, at least, less obviously mistaken.[93] That even those who resist temporal variance for foot faults do not feel about foot faults quite as they do about zone faults—that many of them at least feel the *tug* of temporal variance—requires explanation even if we end up concluding that, all things considered, foot faults should be enforced invariantly.[94] That fact is inexplicable if the argument for temporal variance depends upon the widening of a gap between infraction and penalty and if faults

90. *See supra* note 80 and accompanying text.

91. As will become apparent, it is also the more interesting assumption to indulge because it opens up to us additional possibilities that would not emerge otherwise.

92. *See supra* note 5 and accompanying text.

93. *See, e.g.*, *supra* note 8 and accompanying text.

94. I noted earlier the possibility that temporal variance is warranted by aesthetic considerations: when energy, intensity, and drama have risen, we don't want delays, and we want things to be resolved by excellence, not errors or miscues. While conceding that there is something to this analysis, I contended that it's not the whole story; contemplation of foot faults and zone faults bolsters that suggestion. We prefer the match to end with a winner rather than with an unforced error (including a

aren't penalties for infractions.

I favor our taking widespread intuitions seriously. Doing so invites us to consider whether the analysis supplied thus far furnishes the *only* sound basis for temporal variance. Perhaps it doesn't. Perhaps temporal variance for some power-conferring (or constitutive) rules might be warranted on other (possibly related) grounds.

D. ATHLETIC VIRTUES REVISITED

Recall what I claimed earlier: we want the outcome of athletic contests to depend (insofar as possible) upon the competitors' relative excellence in executing the particular athletic virtues that the sport is centrally designed to showcase, develop, and reward.[95] Call this "the competitive desideratum." It was not a stray or peripheral observation, but rather a linchpin in the case for temporal variance in the enforcement of penalties for incidental infractions in basketball and football. We only ask whether imposition of a penalty unduly affects the competition's outcome when the rule in question does not implicate the core athletic virtues and excellences of the sport in question. *If the rule does implicate the skills that the sport is focally designed to test, then to enforce the rule tends to satisfy rather than frustrate the competitive desideratum.* Thus, we really have two reasons against slack in the calling of balls and strikes. First, as we have seen, calling a nonstrike a "ball" is not a penalty for a rule infraction, but just part of the power-conferring rules of the sport. Second, the ability to throw a strike is just one of the central athletic challenges in baseball. (Having the eye to lay off a nonstrike is a secondary excellence.) So calling strikes as "strikes" and nonstrikes as "balls" promotes our goal that the outcome depend upon the teams' relative excellence in meeting athletic challenges that baseball centrally presents.

What about the rules governing serves in tennis? What are the athletic excellences or challenges involved here? To a first approximation, one of the core challenges is to strike the ball with power and accuracy into a predetermined space. This seems on the right track, though it's clearly not all the way home. Suppose the server were to stand at the net before striking the ball.

---

service fault), but we are *much* less tempted to tolerate temporal variance for zone faults and other unforced errors than for foot faults.

95. I believe that the Supreme Court had something much like this in mind when considering whether the Americans with Disabilities Act required the PGA Tour to accommodate a pro golfer's physical disability by permitting him to ride a golf cart during competition. *See PGA Tour, Inc. v. Martin*, 532 U.S. 661, 700 (2001). In concluding in the affirmative, the majority determined, in effect, that the central athletic challenge the PGA Tour presented was (to a first approximation) the ability to hole a ball by means of striking it with a club, in the fewest number of strokes, while battling fatigue. *Id.* at 690. Justice Scalia, in dissent, charged the majority with presuming to opine on the nature of "Platonic golf." *Id.* at 700 (Scalia, J., dissenting). He was mistaken. The majority was trying to determine what was central or peripheral to the sport of golf *as constituted by actual PGA rules and practices*. That's an interpretive task, and any answer is almost certain to be reasonably contestable. But it is not an inquiry into Platonic essences.

Serving the ball into the service court from there plainly does not satisfy or conform to the athletic challenge that serving in tennis is meant to present. So a refinement is necessary. Perhaps this: the challenge is to strike the ball *into a precisely defined space from a precisely defined distance*.

I'm going to suggest that this is not in fact the best rendition of the athletic challenge that the service rules are meant to embody, and that the challenge is better formulated as the ability to serve the ball *into a precisely defined space from a generally defined distance*. That is, notwithstanding that the formal rules appear to specify both the starting point and the landing space with precision, the "real" or underlying athletic challenge that the rules are designed to codify or facilitate involves a precise target but an imprecise or general launching site. I am tempted to describe the challenge this way: "get the ball *in here* from *around there*." That is surely putting things too loosely, but it conveys my basic claim that precision in the placement of the served ball is of far greater concern to the sport than is precision in the placement of the server's body.

I'll try to make this asymmetry plausible in the remainder of this section. The final two sections will explain why this claim, if true, might support temporal variance for foot faults but not for zone faults.

Let's start with zone faults. Why is the placement of the ball regulated precisely? Why isn't the athletic challenge to get the ball close enough to the service court? Tennis requires that the ball go into the service court because that's the athletic challenge that serving in tennis is designed to serve up. It's how tennis instantiates one of the most commonly tested skills across all of sports: target hitting. And horseshoes and curling notwithstanding, precision is generally part of the nature of targeting—for pitchers, field-goal kickers, and basketball players, no less than for archers and riflemen. To be sure, there is rarely anything essential about the target dimensions. But although the target's contours may be arbitrary, the demand that the competitor hit the target, as it were, and not merely come close, is not arbitrary, for the rule is designed to test and reward that particular class of physical excellences involving accuracy and precision in limb–eye coordination.[96] The rules of tennis require that, for a serve to be valid, the ball must land within the defined service court because that is part of the nature of this particular athletic challenge.

If that's so, why doesn't the same reasoning apply at the front end too? Why isn't precision required with respect to the placement of the server's feet at the

96. Consider some spectacular display of ball-handling artistry in which a soccer player dribbles the ball from deep in his own end up the pitch, deking and dodging defenders all the way until, thirty yards from the goal, he blasts a frozen rope through a tiny window between two defenders that just catches the inner edge of the right goal post and ricochets across the goal mouth before skittering out of bounds. Fans of the sport who are not partisans of either team might well wish that the shot had scored, to reward the player for his dazzling skill. Yet nobody would argue that a goal should be counted. While we might have made the goal a touch larger—indeed, while we might have good reasons for thinking that, all things considered, slightly enlarging the goal would improve the sport—we understand that the challenge is to put the ball into the goal and not to get it "close enough."

moment of striking the ball just as surely as it is required with respect to the location of the ball's landing spot? My claim is *not* that it *couldn't* be. That could be the best understanding of the athletic challenge. I mean only to argue that it needn't be and to suggest that it probably isn't (and of these two arguments, the first is the more interesting and important).[97]

My suggestion will appear more plausible if we start by assuming the opposite. Suppose the athletic challenge that underlies or motivates the rules involves precision at the front end as much as at the back end. What would precision here involve? To start, why should it involve the competitors' *feet* at all? If the challenge were to serve the ball into a specified space from a specified distance, why isn't the relevant distance the distance that the *ball* must travel? Why wouldn't the challenge be better understood to require that the racquet strike the ball behind the vertical plane defined by the baseline? Furthermore, even if the specific distance that should matter is the distance from net to feet, why should we care about the precise location of the feet at the moment the racquet strikes the ball and not be satisfied with specifying the location of the feet at the start of the service motion?

That the feet must remain behind the baseline would make sense were part of the challenge of the serve to strike the ball while keeping one's feet stationary. But that's not the case. Although the written rules of tennis forbid the server from walking or running while serving, they expressly permit "slight movements of the feet."[98] The rules-in-action, moreover, clearly allow much more than that. Given that the sport (sensibly) allows the server to move her feet during what is, after all, a powerful and explosive movement, why should it deem the *precise* location of the feet at the moment of impact of any importance? Although some sports care very much about full bodily control—diving and gymnastics come to mind—tennis is not such a sport. Tennis is a sport of speed and power, hand–eye coordination and overall athleticism. Indeed, the sport is so unconcerned with serving mechanics as even to permit underhand service. It seems much more in keeping with the type of athletic enterprise that is tennis to allow for natural service motions than to micromanage the placement of the feet.[99]

In short, it seems more faithful to the type of sport that tennis is and to the nature of the service to understand the underlying athletic challenge as I have

---

97. Note that the United States Tennis Association recognizes that foot faults and (what I'm calling) zone faults should be treated differently in nonofficiated matches. *See* U.S. TENNIS ASS'N, *supra* note 6, at 13 ("Comment 18.6: . . . . In a non-officiated match, the receiver or the receiver's partner may call foot faults only after all efforts such as warning the server and attempting to locate an official have failed and the foot faulting is so flagrant as to be clearly perceptible from the receiver's side.").

98. INT'L TENNIS FED'N, *supra* note 6 (Rule 18.a).

99. I think that football is different in this regard. Full bodily control is part of the athletic excellence demanded of the wide receiver, which is what made Santonio Holmes's tip-toe touchdown reception in the 2009 Super Bowl such an impressive feat. *See Pittsburgh Steelers WR Santonio Holmes Named Super Bowl MVP*, ESPN (Feb. 2, 2009, 1:24 AM), http://sports.espn.go.com/nfl/playoffs2008/news/story?id=3877889.

described it: to strike the ball into a precisely defined space while starting from a defined place and not traversing unreasonably past the line.

But here's the thing. The rules of tennis don't say that an attempted serve is a fault if the server steps "unreasonably far over the line" or anything of this sort. They say, as we have seen, that the serve is a fault if, during the service motion, "either foot touches . . . the court, including the baseline."[100] Why isn't that the end of it? We needn't speculate as to the nature of the athletic challenge in tennis as presently constituted, the objection goes, for the rules tell us exactly what the challenge is.

To address this complaint lodged on behalf of temporal invariance will require that we introduce one final distinction.

E. TWO MORE KINDS OF RULES

The distinction between duty-imposing and power-conferring rules is one of function. Another, more common, way to categorize rules—it might be clearer to speak now of "norms"—is on the basis of form. This is the division between rules and standards. True, the rule/standard distinction describes something closer to poles on a continuum than to binary possibilities, and any given complex norm can consist of more rule-like and more standard-like pieces cobbled together. Still, the basic difference between the two is fairly well settled: rules turn upon factual predicates that are sharper edged, whereas standards require those who apply them to exercise evaluative judgment. Also well-accepted are the considerations that recommend proceeding with one form or the other. To a first approximation: standards better reflect the genuine justifications that underlie the norm, while rules, because they are quicker and easier to apply, promote second-order considerations in cheap, predictable, stable, uniform, and noncorrupt decision making. As Fred Schauer summarized: "the choice of rule-based decision-making ordinarily entails disabling wise and sensitive decision-makers from making the best decisions in order to disable incompetent or simply wicked decision-makers from making wrong decisions."[101]

Consider the speed limit. There's nothing magical about 65 m.p.h. or 55 m.p.h. or what have you. The real, true, or underlying norm is that people shouldn't drive dangerously fast—or, because a steel contraption weighing one to two tons is dangerous when driven at just about any rate of speed, they shouldn't drive *unduly* dangerously fast. But were we to announce that standard as the norm, we'd end up with a heterogeneity in rates of travel (the unusually timid driving too slow, the overly confident driving too fast) that might itself be

100. *See* U.S. TENNIS ASS'N, *supra* note 6.

101. FREDERICK SCHAUER, PLAYING BY THE RULES: A PHILOSOPHICAL EXAMINATION OF RULE-BASED DECISION-MAKING IN LAW AND IN LIFE 153 (1991); *see also, e.g.*, Larry Alexander & Emily Sherwin, *The Deceptive Nature of Rules,* 142 U. PA. L. REV. 1191, 1198 (1994) (arguing that "a rule works best if it lies").

unsafe. Furthermore, we'd impose additional decisional burdens on police and invite challenge to every enforcement action. Because we think we can identify a numerical speed limit that approximates the true contextual line between safe and unsafe driving tolerably well, albeit imperfectly, we almost always proceed here by rule rather than by standard.[102]

F. TRUE RULES AND RULIFIED STANDARDS

We can now see that the fact that the rule governing foot faults is written in hard-edged terms—a foot fault is defined to exist if either of the server's feet even touches the baseline—is not inconsistent with my claims that the real norm that the rule is designed to implement might be a standard that prohibits servers from going "too far" over the line. Even if the true norm is a standard, it doesn't follow that the formal norm should assume the same shape. Quite the contrary. Because the factors that bear on reasonableness would be debatable in every case, considerations like predictability, certainty, and finality all forcefully favor implementing this norm by means of a rule rather than by means of a standard.

But the part of the service rules that require the ball to land within the service court is different. That, I argued, is part of the underlying athletic challenge. The written criteria of valid service that govern the landing of the ball and the placement of the server's feet are, in both cases, rules rather than standards. But the former is a rule because it captures an aspect of the underlying athletic challenge that is *itself* sharp-edged and rule-like: get the ball *in* the predefined space. The latter is a rule because, although the aspect of the underlying athletic challenge that it captures is standard-like—start from behind the line and don't go unreasonably over it—we have good institutional reasons to codify it in bright-line terms. To coin terms, we might say that that portion of the power-conferring rule of tennis service that requires the serve to land in the service court is a "true rule," whereas that portion of the rule that requires the server not to step on the baseline is a "rulified standard."

Even assuming all this is so, what follows? To start, does it follow that line judges should enforce the rule governing faults as though a foot fault could occur only when the server steps unreasonably far over the line?

Surely not. A rulified standard is a rule, not a standard, and enforcement authorities should generally apply it as such. To routinely pierce the rule and apply the underlying or animating standard would defeat the purposes served by having rulified it in the first place. That must be common ground. It need not,

---

102. We don't always do so, but the rare exceptions tend to bolster the point in text. In the late 1990s, Montana eschewed a rule-like speed limit in favor of a standard that mandated driving "at a rate of speed no greater than is reasonable and proper under the conditions existing at the point of operation . . . . so as not to unduly or unreasonably endanger the life, limb, property, or other rights of a person entitled to the use of the street or highway." MONT. CODE ANN. § 61-8-303(1) (1989); *see also 61-8-303. Speed Restrictions—Basic Rule*, MONT. CODE ANNOTATED 1995, http://data.opi.mt.gov/bills/1995/mca/61/8/61-8-303.htm (last visited May 3, 2011). That proved to be a short-lived experiment. *See* MONT. CODE ANN. § 61-8-303 (1999).

however, be the end of the story.

First, that we must not *routinely* pierce a rulified standard does not mean that we must *never* pierce it. Whether and under what circumstances to disregard the rule's form in favor of its underlying considerations may always be asked with regard to rulified standards. Indeed, that is the most obvious upshot of the distinction between these two types of rules.[103]

Second, it is plausible to suppose that two additional requirements should be satisfied in order to go beneath the surface of a rulified standard: (1) that enforcing the rule as a rule would produce unusually high costs; and (2) that disregarding the rule's form on this occasion would incur low costs on the dimensions, such as predictability and the like, that justified its rulification in the first place.

These two additional conditions, it seems to me, are plausibly satisfied by foot faults in crunch time. The high costs of enforcing the rule as a rule are plain: doing so allows the foot fault to have an undue impact on the match outcome—that is, it thwarts what we have called the "competitive desideratum"—thereby detracting from the participants' satisfaction and the spectators' enjoyment. At the same time, the costs of piercing the rule are very low precisely because the fact of the supposed nonconformity with the rule is hidden from public view. And it's hidden from public view because the Hawk-Eye electronic system that determines whether a ball lands within the lines is not used to judge foot faults. From the perspective of optimal game design, that might be a good thing. In general, rule makers who want to preserve the rule enforcers' option to sometimes apply the standard that animates a rulified standard should arrange things so that noncompliance with the rule isn't apparent. Transparency is not always a virtue.

Of course, even if an ideal system would have (non-expressly) authorized line judges to adjudicate crunch-time foot faults against the underlying standard of reasonableness and not in terms of the nominal rule, that does not fully determine whether Serena Williams's step on the line should have been called. It could be that it was unreasonable or unfair, all things considered—if, for example, her transgression was substantial or repeated. My sense is that it was neither, but I make no strong claims about it. I claim more strongly that Williams's step on the line did not apparently put Clijsters at a competitive disadvantage: the ball landed squarely in the service court and was easily returnable. In sum, if I'm right that the foot fault rule is a rulified standard, not a true rule, that would be a promising (though not conclusive) basis for support-

103. This is not to say, however, that true rules must always be adhered to. Claire Finkelstein has persuasively argued that exceptions to a rule are best conceived as reflections of purposes or principles external to the purposes, principles, or considerations that underlie the rule itself. Claire Oakes Finkelstein, *When the Rule Swallows the Exception*, 19 QUINNIPIAC L. REV. 505, 515 (2000). On this view, true rules and rulified standard alike can be overridden by an exception. My claim is that, for rulified standards but not true rules, we can resist the rule's directive in a second way too—by disregarding its form in favor of the rule's own animating reasons.

ing the McEnrovian intuition: the line judge should have cut Williams some slack.

IV. CONCLUDING THOUGHTS AND ONE SURPRISING LESSON

We started with a puzzle—what might be said in favor of enforcing at least some rules of sports less strictly at crunch time?—and tried to piece together a solution. That solution turned out to be two solutions, or two variants of a single solution. All competitive sports, I have claimed, share a core interest that the outcomes of contests reward competitors' relative excellence in the performance of the sport's fundamental athletic tests. To better serve this interest, each sport has reasons (not decisive or conclusive reasons, but reasons of real weight): (1) not to enforce penalties on infractions when, for contextual reasons, the penalty would be unusually overcompensatory and (2) to sometimes disregard the rule-like form or surface of some norms in favor of the standard that underlies it.

These arguments are tentative and partial, only first steps toward a solution to the puzzle. Whether they ultimately justify the temporally variant enforcement of particular rules of particular sports, all things considered, may strike most jurisprudentially minded readers as of secondary importance. My arguments have expressly drawn on practices and analyses from law and legal theory; moreover, they offer significant promise of returning the favor. For example, they offer insight into the lost chance doctrine in torts, the difference between genuine "jurisdictional rules" and mere claim-processing rules, and the granting of equitable relief in municipal and corporate elections and in appellate litigation. Those promissory notes can't be cashed in this Article, but readers sensitive to the depth and complexity of the philosophical puzzles that arise on the fields of play have reason to suspect that sports will richly repay searching jurisprudential attention.

Still, some readers will hunger for more concrete evidence that the analyses offered here will bear fruit elsewhere. So I'd like to offer one final non-obvious lesson—albeit a lesson for gamewrights, not for legislators. That lesson takes the form of an answer to this question: What's wrong with soccer?

A. A WART ON THE BEAUTIFUL GAME

Not a whole lot. It remains the world's most popular sport by a wide margin,[104] but even many enthusiasts of "the beautiful game"[105] seem increasingly to complain about two defects. Neither is, as Americans might suppose, the low scoring. That, some partisans insist, is part of its beauty. "Soccer is the great team sport because it is a test of team will, and it is a test of team will

104. *See, e.g.*, Doley, *supra* note 16.

105. *See* PELÉ & ROBERT L. FISH, MY LIFE AND THE BEAUTIFUL GAME (1977).

precisely because it is so damned hard to score," rhapsodized one fan:[106]

> You have to run down that field, time and time and time and time again, knowing full well that there's "practically no chance" anything will come of it. Again and again and again. You might have to do it for ninety minutes and get nothing, and then you have to do it again in the next game. It is exhausting, physically and, even more, mentally. But you have to keep doing it, because the moment you stop doing it—the moment anyone on the team starts to think about not doing it—you lose.[107]

If low scoring is a feature of soccer, not a bug, the more oft-noted problems are these: First, there is too much diving ("simulation," in FIFA's more bureaucratic parlance).[108] It disrupts the flow of this continuous-play sport and often provokes unjust penalties. Second, the referees make too many errors. This is partly a consequence of the diving, but it's also a consequence of FIFA's insistence on leaving the officiating to a single referee charged with covering a pitch most often larger than an American football field,[109] and its refusal to introduce any form of instant replay review.[110] Anyone who watched the 2010 World Cup Finals in South Africa—especially, perhaps, followers of the American, English, Mexican, and Irish sides—needs no reminder that the sport could do better by its players and fans.[111]

---

106. David Post, *Americans, Soccer, and Scoring Con't*, THE VOLOKH CONSPIRACY (June 27, 2006, 7:47 AM), http://volokh.com/posts/1151349617.shtml.

107. *Id.*

108. *See* FÉDÉRATION INTERNATIONALE DE FOOTBALL ASS'N, *supra* note 47, at 115 (Law 12—Fouls and Misconduct) (defining simulation as "attempts to deceive the referee by feigning injury or pretending to have been fouled" and providing that it is a yellow-card offense).

109. Calls by soccer line judges are only advisory. *See id.* at 25 (Law 6—The Assistant Referees).

110. Although FIFA has yet to implement instant replay, FIFA president Sepp Blatter has said it will "definitely" consider the technology. *See* Graham Dunbar, *Bad Calls Prompt FIFA To Study High-Tech Ref Help*, USA TODAY (June 29, 2010, 3:46 PM), http://www.usatoday.com/sports/soccer/2010-06-28-933619635_x.htm.

111. But for those who want reminders: The Americans' spectacular comeback against Slovenia was tarnished by referee Koman Coulibaly's still-unexplained disallowance of what would have been the game-winning goal off a direct kick by Landon Donovan; the Three Lions were denied an equalizer against Germany when referee Jorge Larrionda failed to see that a shot by Frank Lampard that hit the underside of the crossbar landed a full two feet inside the goal; Mexico went down 0–1 to Argentina on a goal by Carlos Tevez scored when Tevez was indisputably offside; and Ireland was denied a trip to South Africa when France's Thierry Henry handled to a teammate who booted the ball in for a last-minute goal in qualifiers. For accounts of these games, see Liz Clarke, *Koman Coulibaly's Call Negates U.S. Goal Against Slovenia in World Cup*, WASH. POST (June 19, 2010), http://www.washingtonpost.com/wp-dyn/content/article/2010/06/18/AR2010061805139.html; Elliot C. McLaughlin, *Ireland Outraged After French Handball Nixes World Cup Hopes*, CNN (Nov. 19, 2009), http://articles.cnn.com/2009-11-19/world/henry.handball.world.cup_1_world-cup-irish-national-soccer-team-handball; *Officials Miss Call on Lampard Goal*, ESPN (June 27, 2010), http://soccernet.espn.go.com/world-cup/story/_/id/5333533/ce/us/no-goal-england-refs-miss-first-half-call&cc=5901; *World Cup: Argentina Beats Mexico 3–1*, CBS NEWS (June 27, 2010), http://www.cbsnews.com/stories/2010/06/27/sportsline/main6624435.shtml.

These are problems. I'd like to suggest, however, that soccer harbors a third defect, one that works as a multiplier, exacerbating the first two problems and exacerbated by the fact (not itself a problem) of low scoring. That problem is the red card—in particular that a red card results in ejection of a player for the remainder of the match without allowance given for substitution. Most soccer fans, I think, would be surprised by this diagnosis; I have not seen it voiced before. That it has anything to do with the bulk of this Article might seem doubly surprising. However, I think it follows naturally enough.

A central assumption undergirding the argument that basketball referees should "let 'em play" is that, at least presumptively, the competitive impact of a penalty should bear a stable relationship, over the course of a contest, to the competitive impact of the infraction that the penalty penalizes. We saw, however, that (holding closeness of contest constant) a contest event has a greater impact on the contest outcome the closer it occurs toward the contest's end. Nonenforcement of the penalty at crunch time is an attempt to rectify this imbalance.

This much can be said in favor of soccer's red card: there's no reason to puzzle over whether it should be brandished more reluctantly at crunch time. Unfortunately, that's not because soccer somehow ensures that the red card exerts a constant competitive effect regardless of when it is imposed. To the contrary, the red card exerts a greater competitive effect the *earlier* it is awarded. Because a red card entails the ejection of the offending player along with a ban on his being replaced, it amounts to a requirement that the offending player's team play a man down for the remainder of the match (or at least until the opposing team is red-carded as well). So the more time that remains at the time of infraction, the greater the penalty. In effect, we can imagine that a red card awarded at the fifteenth minute reads "play shorthanded for seventy-five minutes" whereas one awarded for the very same infraction at the eighty-fifth minute reads "play shorthanded for five minutes."[112] The red card thus violates the sensible principle of game design that, presumptively, the same infraction should call forth the same penalty regardless of the time of its occurrence.

112. These are not abstract possibilities, as, once again, play in the 2010 World Cup shows. Consider, for example, the star-crossed Australians. The Socceroos were up 1–0 against Ghana in an opening-round match when defender Harry Kewell was sent off for a hand ball in the twenty-third minute. *See* Rob Smyth, *World Cup 2010: Ghana 1–1 Australia —As it Happened*, GUARDIAN.CO.UK (June 19, 2010, 1:04 PM), http://www.guardian.co.uk/football/2010/jun/19/world-cup-2010-australia-ghana-rob-smyth. It was a harsh and possibly decisive call on a seemingly inadvertent handling. The Black Stars scored on the ensuing penalty kick and held on for a 1–1 draw. *Id.* Days earlier, Uruguay's Nicolas Lodeiro was properly sent off for a reckless challenge against France in the eighty-first minute. *See France, Uruguay Play to Scoreless World Cup Draw*, CBS NEWS (June 11, 2010), http://www.cbsnews.com/stories/2010/06/11/sportsline/main6573038.shtml. Playing defensively, La Celeste could resist Les Bleus for ten minutes, and the game ended in a scoreless tie. *Id.* Australia was ordered to play shorthanded for sixty-seven minutes for Kewell's infraction; Uruguay was disadvantaged for a mere ten minutes for Lodeiro's.

This disparity in the effective magnitude of the red-card sanction should occasion little concern if the optimal penalty for committing a red-card offense (generally, a serious foul, but also such infractions as spitting, handling the ball to deny an obvious goal-scoring opportunity, and committing a professional foul)[113] were ninety minutes of shorthandedness. In that event, the sanction would never be too *high*, and the fact that it would generally be too *low* would be the consequence of something game designers could do nothing about.[114] It would be much like the problem that the criminal justice system faces when encountering a seventy-year-old offender who commits a crime for which we believe the optimal punishment is forty years imprisonment. Being unable to ensure that the convict lives to 110, we just shrug at the supposed fact that he will be getting off lightly.

As far as I can tell, however, we have no reason at all to believe that the optimal sanction for the sort of infraction that calls forth a red card is being shorthanded for ninety minutes. To be sure, what would be an optimal period of shorthandedness is extraordinarily difficult to determine, but the basic parameters for thinking through the problem are clear enough.

A red card is awarded for a serious offense. The offending team should incur a significant penalty—the offense should significantly affect the team's prospects for victory. At the same time, the penalty should not be virtually outcome determinative—all the more so given the prospect (exacerbated, recall, by the prevalence of diving, by the presence of a lone referee, and by the absence of replay) that some whistles for a red card will be in error. Nobody, for example, would seriously entertain a proposal to replace the penalty of ejection with the award of two goals to the opposing team. Given soccer's low average scores and margins of victory,[115] a sanction of such magnitude would threaten to convert the sport into an extended exercise in penalty avoidance. Similarly, we might expect that sending off a player in, say, the tenth minute is apt to have such a significant impact on game outcome as to contravene the competitive desideratum.

---

113. *See* FÉDÉRATION INTERNATIONALE DE FOOTBALL ASS'N, *supra* note 47, at 35 (Law 12—Fouls and Misconduct).

114. This isn't strictly true. Presumably, some number of minutes of being down two players would have the same competitive impact of being down one player for ninety minutes. If system designers really wanted an offending team to suffer the competitive disadvantage that corresponds to being one player down for ninety minutes, they could, for example, specify that a red card awarded in the second half required the ejection of the offender and one teammate. But this is a quibble.

115. Reliable relevant statistics are hard to come by, in part because of the multiplicity of professional leagues and the variety of international competitions. To get a flavor of the magnitudes involved, though, note that during the past five seasons English Primer League teams scored an average of 1.52 goals at home games and 1.07 goals at away games and that the average score in the sixty-four games of the 2010 World Cup in South Africa (not including penalty shoot outs) was 1.75–0.52 (these results were calculated, respectively, from *England Football Results Betting Odds*, FOOTBALL-DATA.CO.UK (Jan. 26, 2011), http://www.football-data.co.uk/englandm.php, and *The Matches of 2010 FIFA World Cup*, FIFA, http://www.fifa.com/worldcup/matches/index.html (last visited Jan. 29, 2011)).

B. A QUICK FIX

If this is right, what's to be done?

One possibility, of course, is temporal variance—but in reverse. Soccer referees should be more reluctant to administer red cards early in matches than late. In fact, this may be current practice. One study of all the games played in the German first Bundesliga over forty-one seasons determined that more than forty percent of the first red cards issued in a match were issued in the game's final twenty minutes whereas less than seven percent occurred in the first twenty-three minutes.[116]

There's a second option: to unlink the penalty of ejection from the penalty of shorthandedness. Soccer already decouples the consequences of a red card for the player involved from the consequences for his team: the player is sent off for the remainder of the match and is disqualified for the next game too,[117] but the team plays shorthanded only for the remainder of the current game, not for the next. Soccer's governing bodies should consider taking this decoupling further. That the offending player may not return does not entail that his team should play shorthanded for the rest of the contest regardless of when the foul occurred. Many sports—including hockey, team handball, and indoor soccer—allow a team to substitute for an ejected player after some period of penalty time.[118] Perhaps soccer should follow their lead. To require a team to play

---

116. Michael Bar-Eli et al., *Consequences of Players' Dismissal in Professional Soccer: A Crisis-Related Analysis of Group-Size Effects*, 24 J. SPORTS SCI. 1083, 1087 (2006). Of course, it is possible, additionally or alternatively, that players behave differently early and late in games precisely because the impact of a red card varies depending on when it is issued.

117. This is itself a controversial feature of dismissal because it ensures that *two* teams—the offending team's current opponent and its next opponent—benefit from the commission of a red-card offense by a third team. Because it might be essential to the current opponent that the offending team defeat its next opponent (as when the first opponent and the second are vying for a sole remaining chance to advance), this feature of the red card system seems to add insult to injury. (I thank Stephen Weatherill for first drawing this objection to my attention.)

118. *See, e.g.*, NAT'L HOCKEY LEAGUE, OFFICIAL RULES 2010–11, at 30 (Rule 20.3 Substitution) ("When a player has been assessed a major penalty and has been removed from the game or is injured, the offending team does not have to place a substitute player on the penalty bench immediately, but must do so at a stoppage of play prior to the expiration of the major penalty."); *Indoor Soccer 6 v 6 Laws of the Game*, OXNARD INDOOR SOCCER, http://oxnardindoorsoccer.com/Rules_and_Forms.html (last visited May 3, 2011) (providing that when a player is ejected "the team can replace the player so they do not play short"). In team handball a "suspension" requires the offending player to sit out for two minutes, during which time his team plays shorthanded, whereas a "disqualification" results in the ejection of the offending player for the remainder of the contest, but only a two- or four-minute period of shorthandedness for this team. *See* INT'L HANDBALL FED'N, RULES OF THE GAME 49, 51 (2010) (Rules 16:3, 16:6). Until a recent revision, handball also authorized a penalty that, like the present red card in soccer, resulted in dismissal of the offender for the remainder of the game without right of substitution. Called an "exclusion," it applied to a "forceful and deliberate attack against the body of another person." INT'L HANDBALL FED'N, RULES OF THE GAME 33, 54 (2005) (Rules 8:7, 16:9). My suggestion is that soccer should not maintain a red card system that is the mere equivalent to the old "exclusion" in handball. That is, I propose that it change the current red card to be something like "disqualification" in handball. Whether soccer should preserve its very strong medicine *in addition* to the somewhat softer penalty—perhaps it could have a crimson card and a burgundy card—is a separate question. Of course, this question again implicates the lumping/splitting debate. *See supra* note 73.

shorthanded for nearly a full game is draconian even when the offense really warranted dismissal. But it's heartbreaking when—as happens disappointingly often in this otherwise beautiful game—the red card should never have been issued.

To be sure, figuring out what would be an appropriate period of shorthandedness would prove challenging. The task involves, at a minimum, both an evaluative question (what should be the approximate cost of a red-card offense, measured in the coin of impact on probability of victory or of impact on expected goal differential) and an empirical one (what is the expected fractional goal differential per minute of shorthandedness). Presumably, econometricians would be beyond eager to address this second question were FIFA to put it on the table.[119] But, admittedly, neither question is a gimme.

That's okay with me, for I don't purport to be confident regarding how to improve soccer. I claim only that it's hard to believe that the current system that makes the competitive impact of a red card so radically dependent on its time of issuance truly dominates the alternatives and, therefore, that further investigation is warranted. More to the point: that we should think harder about soccer's red-card system is only one among the many and diverse lessons to be learned by reflecting on Serena Williams, Don Larsen, and the law.

---

119. This is not an easy nut to crack. Several recent papers have reached conflicting conclusions on the simpler question of whether—as the saying "ten do it better" would have it—being shorthanded actually improves team performance by forcing an increase in individual effort and team play. *See* Bar-Eli et al., *supra* note 116; Fiona Carmichael & Dennis Thomas, *Home-Field Effect and Team Performance: Evidence from English Premiership Football*, 6 J. SPORTS ECON. 264, 277–78 (2005) (finding that red cards are less costly for away teams); G. Ridder et al., *Down to Ten: Estimating the Effect of a Red Card in Soccer,* 89 J. AM. STAT. ASS'N 1124, 1126–27 (1994) (finding that an early red card increases the chance of victory for the other team); Marco Caliendo & Dubravko Radic, *Ten Do It Better, Do They? An Empirical Analysis of an Old Football Myth* 1, 13 (Institute for the Study of Labor Discussion Paper No. 2158, 2006), *available at* http://ssrn.com/abstract=908250 (finding that scores, for either team, were not influenced by expulsion and that the idea that "ten do it better" remains a "myth"); Mario Mechtel et al., *Red Cards: Not Such Bad News for Penalized Guest Teams* 15 (Working Paper, 2010), *available at* http://ssrn.com/abstract=1571867 (finding that the send-off for a guest team after seventy minutes led to worse score for home team).