FROM: THE VOLOKH CONSPIRACY

"LET 'EM PLAY"

Mitch Berman[†]

OVERVIEW

Many thanks to Eugene for inviting me to discuss my just-published paper "Let 'em Play": A Study in the Jurisprudence of Sport,[1] in this forum. I'm grateful for the opportunity and look forward to your comments.

Recall the women's semifinal of the 2009 U.S. Open, pitting Serena Williams against Kim Clijsters. Having lost the first set, Williams was serving to Clijsters at 5–6 in the second. Down 15–30, Williams's first serve was wide. On Williams's second service, the line judge called a foot fault, putting her down double-match point.

Williams exploded at the call, shouting at and threatening the lineswoman. Because Williams had earlier committed a code violation for racket abuse, this second code violation called forth a mandatory one-point penalty. That gave the match to Clijsters.

Williams's outburst was indefensible. But put that aside and focus on the fault. CBS color commentator John McEnroe remarked at the time: "you don't call that there." His point was not that the call was factually mistaken, but that it was inappropriate at that point in the match even if factually correct: the lineswoman should have cut Williams a little slack. Many observers agreed. As another former tour professional put it,[2] a foot fault "is something you just

---

[†] Richard Dale Endowed Chair in Law, University of Texas School of Law. Original at volokh.com/author/mitchberman/ (July 18-22, 2011; vis. Oct. 1, 2011).

[1] 99 Geo. L.J. 1325 (2011).

[2] Michael Wilbon, *A Call and a Response That Cannot Be Defended*, Wash. Post, Sept. 14, 2009, www.washingtonpost.com/wp-dyn/content/article/2009/09/13/AR2009091302533.html.

don't call – not at that juncture of the match."

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. 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." On this view, which probably resonates with a common understanding of "the rule of law," sports rules should be enforced with resolute temporal invariance.

Perhaps McEnroe was wrong about Williams's foot fault. But the premise of the *Times* letter – that participants and fans of any other sport would reject temporal variance decisively – is demonstrably false. 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."

Regardless of just how general this supposed agreement is, many NBA fans would affirm both that contact that would ordinarily constitute a foul is frequently not called during the critical last few possessions of a close contest and that that is how it should be. So insistence on rigid temporal invariance requires argument not just assertion.

However, advocates of temporal variance shouldn't be smug either. For while the negative import of temporal variance is clear – the *denial* of categorical temporal *in*variance – its positive import is not. Surely those who believe that Williams should not have been called for a fault implicitly 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 would contend that pitchers should be awarded extra inches 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 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.

"Let 'em Play" is an attempt to think through this problem. My goal is *not* to establish whether and in what respects temporal variance is optimal, all things considered, for any given sport. That's too darn hard.

My goal at this early stage is merely to figure out whether "sense can be made" of such a practice. Instead of trying to determine conclusively just what optimal practices should be, I aim only to explain why temporally variant rule enforcement might be sensible – what can plausibly be said for it.

Furthermore, investigating temporal variance in sport is only the paper's surface agenda.

While econometricians are busily tackling sport, and while philosophers of sport occasionally draw on legal philosophy (in addition to, e.g., aesthetics, ethics, and metaphysics), legal theorists have paid sports only passing attention. Most jurisprudential appeals to sports and games have been ad hoc, and most legal writing on sports that does not pertain to sports law is intended more to entertain than to edify.[3]

The lack of sustained jurisprudential attention to games, and sports in particular, should surprise, for sports leagues constitute distinct legal systems. This is superficially apparent to non-Americans. While baseball, football, and basketball are governed by official "rule books," the most popular global team sports like soccer, cricket, and rugby are all formally governed by "laws," not "rules." More substantively, sports systems exhibit such essential institutional features as legislatures, adjudicators, and the union of primary and secondary rules.

Accordingly, my grander ambition is to help spur the growth of the jurisprudence of sport as a field worthy of more systematic attention by legal theorists and comparativists. In a sense, *"Let 'em*

[3] *Aside: The Common Law Origins of the Infield Fly Rule*, http://www.pennumbra.com/issues/pdfs/157-1/Infield_Fly_Rule.pdf, and 123 U. Pa. L. Rev, 1474 (1975).

Play" does double duty as a manifesto for an enlarged program of jurisprudential inquiry.

Importantly, it's not just that (municipal) legal systems and sports systems confront similar challenges. For several reasons, jurisprudential attention to sports is particularly likely to contribute to our understanding of phenomena and dynamics shared in common.

First, because sports' rules and practices have long been thought unworthy of serious philosophical investigation, even low-hanging fruit has yet to be harvested. Second, sports supply vastly many examples for the generation and testing of hypotheses. And third, our judgments and intuitions about certain practices – such as, to take the present topic, the propriety 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.

For all these reasons, sporting systems, though rarely explored with seriousness by legal theorists and comparative lawyers, comprise a worthy object of legal-theoretical study.

Here's my plan for the remainder of the week. Tomorrow, I will summarize my prima facie case for temporally variant enforcement of non-shooting fouls in basketball and, by extension, of similar violations in other sports. In a nutshell, that argument depends upon a growing gap between the competitive cost of the infraction and the cost of the sanction imposed for the infraction.

On Wednesday, I will explain why the argument that might explain and justify temporally variant enforcement of fouls in sports like basketball, hockey, and football most likely does not cover the rules governing faults in tennis. On Thursday I will propose a different account that might fill that need – one that draws on what I think are novel observations about the hoary rules/standards distinction.

On Friday, I will advance a modest proposal for improving the world's most popular sport.

Tags: basketball, discretion, foul, jurisprudence, penalty, sports, tennis. 45 Comments.

A FIRST SOLUTION

Although the Serena Williams episode provoked my interest in the puzzle of temporal variance, I'll start not with tennis, but with other sports in which a practice of temporal variance might seem 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!" or "Swallow the whistle!"

Though the plea is familiar, its rationale is obscure. To be sure, the tighter the rules are enforced, the less physical contact there will be. And observers may reasonably disagree about the level of physicality that makes a sport the best it can be.

But however a league might answer that question, it is not self-evident why the optimal degree of laxity should differ in crunch time during an NBA game relative to ordinary time, or throughout the NHL playoffs relative to the regular season. 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."

Still, basketball remains a good place to start. I doubt that many tennis fans are justifiably confident that tennis officials do (or don't) allow players a little more foot faulting toward the end of close matches than earlier. Maybe they do (or don't), but foot faults just aren't called enough to permit those without intimate knowledge of the sport to be sure what the enforcement patterns are.

Basketball is different. That basketball referees respect some measure of temporal variance seems clear to many hoops fans. Maybe that's because the case for temporal variance in basketball is unusually clear. (Or maybe not.) If we can explain and justify slack in the calling of basketball fouls, we might be better able to assess whether temporal variance makes sense elsewhere too.

One rationale for temporal variance invokes essentially aesthetic considerations: the referee's whistle disrupts play, thereby reducing spectators' enjoyment of the action. And while disruption of play almost always incurs an aesthetic cost, disruption during crunch time is especially costly (aesthetically speaking) given heightened

dramatic tension.

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.

For example, it appears to me (and not only to me[4]) that football officials are often more reluctant to call defensive pass interference during crunch time even though an incompletion stops play just like a penalty flag. Because an aesthetic or dramatic preference that play continue unabated wouldn't seem to explain or justify temporal variance everywhere it appears, 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, we have reason to look for an alternative account too.

A second answer, recently advanced by Chicago economist Tobias Moskowitz and SI columnist L. Jon Wertheim in their book Scorecasting,[5] depends entirely on the omission bias. By relying entirely on a cognitive bias, however, the authors all but ensure that, even insofar as their account might help explain temporal variance, it is unlikely to *justify* it.

The alternative account I offer runs as follows:

(1) In the main, a sanction imposed for an infraction has a greater expected impact on contest outcome (against the rule-violator) than does the infraction itself (in the violator's favor). This must be so for the sanction to serve a deterrent function in addition to a restitutionary one.

(2) The expected impact of all outcome-affecting contest events – e.g., scores, base hits, yardage gains, infractions, penalties, etc. – are not constant, but context-variant. To start: the closer the contest, the greater the impact. The variance that matters for my purposes, however, is temporal: when the contest is close (and holding

[4] Peter King, *Monday Morning Quarterback*, Sports Illustrated, Nov. 23, 2009, sportsillustrated.cnn.com/2009/writers/peter_king/11/22/Week11/3.html.

[5] *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won* (2011), scorecasting.com.

the closeness of the contest constant), the expected impact of outcome-affecting events varies in inverse proportion to the distance remaining to contest's completion.

For example, touchdowns and baskets, 15-yard penalties and free throw opportunities, all have greater impact on the expected outcome when occurring 2 minutes before the end of a then-tied game than when they occur 2 minutes from the start. (I expect pushback here, and look forward to debates in the comments.)

(3) From (1) and (2) it follows that the absolute magnitude of the gap between the competitive impact of the infraction (say, a non-shooting foul) and the competitive impact of the penalty imposed for the infraction (say, the award of free throws) is significantly greater in crunch time during close games than earlier in the same contest. The penalty becomes more overcompensatory in absolute terms.

(It does *not* become more overcompensatory in *relative* terms, which is why some of yesterday's posters rightly observed that if the stakes become higher for the competitor who would wish to invoke temporal variance, they become higher for their opponents too.)

(4) It is a general principle of competitive sport that athletic contests go better insofar as their outcomes reflect the competitors' relative excellence in executing the particular athletic virtues that the sport is centrally designed to showcase and reward. (This is a first cut; no doubt my proposed principle could be profitably refined further.) This is why we prefer to reduce the impact of luck on outcomes (e.g., we generally want playing surfaces to be regular thus reducing unpredictable bounces).

It is also why almost everybody agreed, in Casey Martin's lawsuit against the PGA,[6] that if (as the Supreme Court majority essentially concluded, but as the dissent denied) the central athletic challenge the PGA Tour presented was the ability to hole a ball by means of striking it with a club, in the fewest number of strokes, while battling fatigue, then golf is less good – it exemplifies a core value of sport less well – if it requires competitive golfers to walk

[6] *PGA Tour, Inc. v. Martin*, 532 U.S. 661 (2001), at http://www.law.cornell.edu/supct/html/00-24.ZS.html.

the course even when it is extraordinarily difficult for them to do so and when they are greatly fatigued without walking.

(5) From (3) and (4) we have a reason (not a conclusive reason) to enforce restrictions on minor or incidental contact less strictly toward the end of close contests if – as is contestable but surely plausible – the ability to refrain from minor bodily contact with opponents is a peripheral athletic virtue in basketball as we know it. If this is so, then a penalty of nominally constant magnitude that it is optimal to impose early in a contest may become suboptimal later in that same contest.

To be clear: I do not claim that the excellence of avoiding minor contact is something that no sport could wish most to valorize. My argument for temporal variance in basketball is explicitly contingent on its being the case that this particular excellence does not rank so highly among the excellences that basketball wishes to feature and encourage. Whether this is so is an interpretive question.

That's my proposed pro tanto argument for temporally variant enforcement of non-shooting fouls in basketball. The argument extends to similar fouls in sports like football and hockey. At bottom, it's based on an aversion to the awarding of windfall remedies disproportionate to the harm suffered. That's a principle the law frequently endorses – from the harmless error rule to contract law's material breach doctrine.

83 Comments.

OF CONSECUTIVE AND NEGATIVE RULES

At first blush, we might suppose that the analysis I provided yesterday applies, mutatis mutandis, to foot faults in tennis and therefore that tennis officials should call foot faults less strictly at crunch time. But this conclusion would be premature. It could be that foot faults in tennis differ from fouls and similar infractions in basketball, football and comparable sports in ways that make a difference.

I'll explain today why I believe that foot faults *do* differ in a way that matters. Tomorrow I'll argue that temporal variance in their enforcement might nonetheless be defensible on alternate

grounds. This afternoon I will respond to some of the many excellent comments already posted by VC readers.

The analysis I presented yesterday for temporal variance in the enforcement of penalties for fouls like those committed in basketball depended upon the claim 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 both 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 some cases. Take balls and strikes in baseball. 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.

That not all undesired consequences that attach to nonconformity with the dictates of a rule are sanctions imposed for infractions was a central claim upon which Hart relied when critiquing the Austinian command theory of law.

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 "if you wish to do this, this is the way to do it." 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 is designed to maintain."

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 will be invalid otherwise. The power-conferring/duty-

imposing distinction is, at a minimum, a close cousin to another distinction between rule types made famous by John Searle: the distinction between constitutive and regulative rules.

The Hartian analysis of power-conferring rules helps to explain why balls and strikes in baseball feel very different from the infractions I have discussed in basketball. In the case of the latter, we can sensibly ask both whether some type of contact ought to be proscribed (thus denominated as a "foul"), and, in addition, whether, if so, 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 its being outside the strike zone is that it is a ball. While 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), or whether any number of balls should result in the award of a base, it seems nonsense to ask whether a pitch's being a ball is too high a price for its having narrowly missed the strike zone: that the pitch was a ball is just what it means for its not having been a strike.

In short, balls and strikes are not proper candidates for temporal variance on the analysis I sketched yesterday 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 constitutive rules, or something of this sort).

If this is right, the question becomes whether the rules governing foot faults in tennis are power-conferring (or constitutive) as opposed to duty-imposing (or regulative). For want of space, I'll just assert that the former construal seems significantly more plausible. 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 that these are three components of the rule that defines a valid serve. A failure on any of these three grounds is just a failure to perfect the power conferred upon the server; none is a violation or an infraction.

Let's suppose that's correct. Even if so, here's the puzzling thing. If foot faults, just like ordinary "zone" faults (i.e., the failure to serve the ball into the service box), are 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 surely as are zone faults. But widespread intuitions are more equivocal.

I have not run across anybody who is tempted by 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 some folks (McEnroe, for example) 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 the least, more plausible, less obviously mistaken. The fact that even those who resist temporal variance for foot faults do not feel about foot faults quite as they do about zone faults – the fact 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. That fact is inexplicable if the argument for temporal variance depends upon the widening of a gap between infraction and penalty and if faults 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. That's my topic for tomorrow.

36 Comments.

SOME RESPONSES TO COMMENTS

First, let me thank the many readers who have commented these past few days. I did not know what to expect when I accepted Eugene's invitation to blog about my article, and have been impressed by, and grateful for, the number and incisiveness of the comments. Unfortunately, there have been too many to permit me to respond in a systematic manner, let alone in a comprehensive one. So here are a mess of somewhat random reactions.

1. I've agreed with many of the posts, and have been gratified to see that many readers anticipated arguments to come.

For example, Assistant Village Idiot observed that my analysis "would suggest that a possible strategy would be to reduce the penalty late in the game but call it more closely. I don't know if that would actually play out well, however."

Agreed on both counts. See p.1349 n.73 of my article for some remarks on just this score.

Soronel Haetir remarked on Tuesday: "I can see some argument for allowing more contact later in a game (an argument I don't particularly agree with), but I don't see any reason whatsoever for relaxing the basic rules of ball possession." I hope that this morning's post revealed my full agreement that the argument I offered on Tuesday would not support relaxing "the basic rules of ball possession." Those are constitutive rules.

Justin agreed with Tuesday's analysis but added: "except for fouling out in basketball and red cards in soccer. Two fouls called on a key player in the first 5 minutes of a basketball game can change the entire contest. And a soccer team playing 80 minutes while a man down is almost certain to lose."

So true. Wait for Friday. Incidentally, Friday's post will simplify matters by ignoring Visitor Again's observation that soccer refs might already respect temporal variance in the issuance of red cards. This is addressed in the article at p.1368 & n.116.

Guy and I seem to be on the same page. I agree with his observation on the regulative/constitutive distinction that "the distinction is

less something that can be derived by objective observation of the law in operation, but more by how people understand the law and what its purposes are." He then added: "the most obvious distinction between foot faults and zone faults is that most people think of the game as being a test of skill with respect to hitting the ball, not where you place your feet. Foot faults only exist because the game needs to prescribe a spot for you to serve from, but minor variations in the rule are unlikely to change the difficulty of performing a proper serve. Rigid adherence to the rule is probably thought of as more penal by the audience than rigid adherence to zone fault rules because the game is 'testing' your ability to hit the ball precisely to serve to a particular spot, but it isn't 'testing' your skill at putting your foot close to a line without going over."

Yep, that will a core piece of tomorrow's argument. Incidentally, Justin agreed with Guy, but added: "Unfortunately, I think one of the problems with your analysis is that you are looking at it through a legal philosophy prism when the answer you are looking for is an anthropological one." This puzzled me. Anthropology and philosophy needn't be at odds. I understand my philosophical analysis to point out which anthropological facts are relevant, in what ways, and why. Perhaps Justin might further explain why he thought his observation showed a problem with my analysis (or with Guy's?).

Lastly, I think Martinned is right, as against both Noah and Gentleman Farmer, that the relative distinction is not objective/subjective.

2. The problem of time-sensitive impact.

I received fewer challenges than I anticipated to my claim that outcome-affecting events have greater impact the later they occur in a close contest, holding closeness of contest contest. I believe only Bruce Boyden and Tom Swift objected.

Here are a few additional thoughts on the matter. I think almost all of us feel comfortable saying things like Team A has a .X probability of winning this game. We believe, for example, that the U.S. women's soccer team had a pretty high probability of victory immediately after Abby Wambach's goal. We believe that the team's

probability of victory was lower once Japan equalized. Almost all probability theorists believe that such statements are meaningful and that they must be some type of subjective probabilities. (The objective probability of a U.S. victory was, at all times, 0.)

If we then believe that events can affect outcome-probabilities, we must be comfortable assessing these things in terms of subjective probability. And once we're in subjective probability land, my claim that late events change the probabilities more than early events do is quite sound as a generalization, though there can be exceptions. (See, e.g., p. 1350 n.74.) Given all this, I'd need to hear more from Bruce Boyden regarding why he believes that the perspective of an omniscient observer supplies the "more relevant comparison."

Tom Swift is surely right in one sense that "points count the same at the beginning of a game as they do in the last 2 minutes." They count the same in terms of nominal additions to the score. But they don't count the same in terms of changes to probability of winning so long as the relevant probability is subjective – which, I've just said, it must be so long as we continue to make claims about probability less than 1 and greater than 0.

3. Miscellaneous thoughts.

Many of the remaining posts raised ideas that might not be strictly germane to my arguments thus far, but which I found interesting enough to merit some reaction.

tbaugh wrote:

> I've never understood how an official not calling a violation late in the game is "letting the players and not the officials decide the game." A non-call of a violation is an official influencing the game, perhaps decisively. I think the comment from James about uncertainly in the determination of an infraction is a good one, however, particularly in basketball. Perhaps some "temporal variance" is justifed in terms of the degree of certainty the official should have in making a late call (I've done a litte refereeing, and I'd say it's kind of a "felt" thing rather than a conscious decision).

I wonder whether the ideas in this post are in tension. Temporal

variance in degree of certainty (actually, the NBA has a rule about this!) would make sense if the costs of false positives and false negatives differ toward contest's end. But tbaugh seems to deny that. I happen to agree that temporal variance in the standard of proof makes sense. But the judgment that a false positive is worse than a false negative is (and must be, I think) parasitic on the supposition that the sanction and the penalty are differently costly as measured against the competitive desideratum. (Incidentally, James's different argument for why uncertainty might lead to temporal variance seems largely dependent upon omission bias.)

duffy pratt observed that "Baseball has a different time element than other games" and asked for examples "where this idea of "temporal variance" would apply in baseball?"

I'm disposed to think that baseball has few good examples not because it has a different time element (see 1336 n.32) but because it has few duty-imposing/regulative rules and many power-conferring/constitutive ones. I do think that balks provide a good potential example, though.

Ossus recalled

> baseball announcers advocating a form of situational (if not strictly temporal) variance with balls and strikes. For example, on 0–2 counts when the batter takes a close pitch, I have heard announcers talk about how the umpire either should have (when they call a third strike) or did (when they call a ball) take the situation into account. The implication is obviously that the penalty for a called strike to the batter is much greater than the penalty of a called ball to the pitcher, so I think this can actually fit into your analysis whereas you claim that it does not.

The analysis in a book I mentioned earlier, Scorecasting, reveals that umpires do take the situation into account in must this way. I am disposed to believe that they ought not to. More interestingly, as some commentators observed previously, Steven Jay Gould thought that home plate umpire Babe Pinelli rightly gave Don Larsen a few extra inches on his last called strike to end his perfect game in the 1956 World Series. I differ with Gould here. (See pp. 1352-54)

Lastly, Byomtov opined that "calling a pitch a ball is a penalty, or at least can be seen as one. If we say the idea of the game is for the batter to try to hit the ball, etc., then there needs to be a rule requiring the pitcher to throw it where the batter actually can reach it. The penalty for violating the rule four times is a walk." I think that's an interesting analysis. Balls could have arisen as Byomtov conjectures and still count as constitutive rules today. I'll think more about this.

Byomtov also remarked, presumably tongue-in-cheek, that he "wouldn't be surprised if the rule was established – by Abner Doubleday no doubt – precisely for this purpose, though of course it turned out that it often makes sense to violate it and suffer the penalty."

Interestingly, early baseball had no bases on balls. There were balls, but no number of balls resulted in a free pass to first. I believe that bases-on-balls were introduced in 1879. At that time, though, a pitcher had 9 balls for a walk. The current rule that awards a walk on 4 balls was introduced ten years later.

That's it for now. See you tomorrow.

24 Comments.

OF RULES AND STANDARDS

Recall Tuesday's contention: Competitive sports go better, all else equal, insofar as contest outcomes reflect the competitors' relative excellence in executing the particular athletic virtues that the sport is centrally designed to showcase, develop and reward. Call this "the competitive desideratum." If something like this is so, then we should identify the athletic challenges that the rules governing tennis serves are designed to hone and test.

To a first approximation, the challenge is to strike the ball with power and accuracy into a specified space. Yet serving while standing at the net would not conform to the athletic challenge that tennis service 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*.

Notice that if this is the best understanding of the athletic challenge presented by serving in tennis, then temporally variant enforcement of foot faults would not serve the competitive desideratum. If it's constitutive of a core athletic challenge in tennis to hit the serve without touching the line, then to forgive a server's having stepped on the line would frustrate that athletic ideal and would contravene the competitive desideratum.

But perhaps that is not quite the athletic challenge that the service rules embody. Perhaps 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 specify both the starting point and the landing space with precision, the underlying athletic challenge that the rules codify involves a precise target but a general launching site.

I am tempted to describe the challenge this way: "get the ball *in here* from *around there*." That puts things too loosely, but it conveys that the sport might care more about precision in the placement of the served ball than precision in the placement of the server's body.

Arguments could be mustered to bolster this interpretation of the core athletic challenge in serving. But I concede that it's debatable. Let's move on because my jurisprudential ambitions are served by exploring what might follow if this is the better conception of the athletic challenge; it's not essential to establish that this is the better interpretation of tennis.

Importantly, that the foot fault rule is *written* in hard-edged terms does not disprove that the real norm the rule implements is a standard that prohibits servers from going "too far" over the line, or that prohibits "unreasonable" encroachments. Even if the true norm is a standard, it doesn't follow that the formal norm should assume the same shape.

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. This is Rules vs. Standards 101.

In short, I am suggesting a critical asymmetry. The written crite-

ria 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 they are formulated as rules for different reasons.

The former is a rule because it reflects an aspect of the underlying athletic challenge that is *itself* sharp-edged and rule-like: get the ball in the pre-defined space. Tennis rules require that the ball go into the service court because that's the nature of the challenge of serving. It is how tennis instantiates one of the most commonly tested skills across all of sports: target-hitting. Horseshoes and curling notwithstanding, precision is generally part of the nature of targeting.

Although a target's contours may be arbitrary, the demand that competitors hit the target and not merely come close is not arbitrary, for the rule is designed to test and reward that particular class of physical excellences (needed by, e.g., archers and riflemen) involving accuracy and precision in limb-eye coordination. The rules of tennis require that, for a serve to be valid, the ball must land within the defined service court because that is the nature of this particular athletic challenge.

In contrast, the formal norm governing foot placement is rule-like not standard-like, I suggest, because, although the aspect of the underlying athletic challenge that it captures is standard-like (start behind the line and don't go unreasonably over it), we have good institutional reasons to codify it in bright-line fashion.

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." It is often thought that norms are standard-like in what we might call their "natural" state, and that they become rules, when they do, in response to institutional pressures. I am suggesting that this is true of some norms but not all. Some of the rules we come across are rules naturally.

Granting me all this, 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? No. A

rulified standard is, after rulification, a rule, not a standard. To routinely pierce the rule and apply the underlying or animating standard would defeat the purposes served by having rulified it.

But that we must not *routinely* pierce a rulified standard does not mean that we must never pierce it. Whether to disregard the rule's form in favor of its underlying considerations is always at least askable with regard to rulified standards. That is a central upshot of the distinction between rulified standards and true rules.

At least two additional requirements must be satisfied to pierce 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 warranted its rulification.

These two additional conditions are probably satisfied by foot faults in crunch time. Enforcing the rule as a rule is costly because doing so allows the foot fault to unduly impact the match outcome. That is, it undermines the "competitive desideratum." And the costs of piercing the rule are low because nonconformity with the rule is hidden, given that tennis does not employ its Hawk-Eye electronic system to judge foot faults.

From the perspective of optimal game design, that might be a good thing. Rule makers who want to preserve rule-enforcers' discretion to sometimes apply the standard that animates a rulified standard should arrange things so that non-compliance with the rule isn't apparent. Transparency is not always a virtue.

Of course, even if the ethos of tennis should permit line judges to assess crunch-time foot faults against the underlying standard of reasonableness, not against the nominal rule, that does not fully resolve the Serena Williams case. Her foot fault would have run afoul even of the standard if, for example, her transgression was substantial or repeated. I think it wasn't, but needn't argue about that here.

In sum, my analysis is doubly contingent: if the foot fault rule is a rulified standard not a true rule, and if Williams complied with the underlying standard-like norm governing service, we'd have promising support for McEnroe's contention: the line judge should have cut Williams some slack.

CONCLUDING THOUGHTS

I started on Monday 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 develop 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 further this interest, each sport has reasons – weighty but not decisive – (1) not to enforce penalties on infractions when, for contextual reasons, the penalty would be unusually over-compensatory, 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. But whether they ultimately justify the temporally variant enforcement of particular rules of particular sports, all things considered, is not greatly important to me. Think of this study as a search for what Robert Nozick called a philosophical explanation: not a defense of the thesis that temporal variance in sports is optimal, but an account of how that could be.

Philosophical explanations are not always the right goal. Often we want to know what some agent should do. In this case, however, I'm satisfied to identify factors and analytical devices that might prove useful for theoretical projects across reaches of law and sports.

For example, the analyses here might helpfully illuminate the lost chance doctrine in torts; the granting of equitable relief, near contest's end, from rules governing municipal and corporate elections, or appellate litigation; the difference between genuine "jurisdictional rules" and mere claim-processing rules; and possibly much else.

Those are just promissory notes at this point. So I'll conclude by offering one final non-obvious lesson – albeit one for gamewrights, not for legislators or judges. It concerns soccer.

Here are two much-noted problems with the beautiful game: there is too much diving, and refs make too many errors. The latter

is partly a consequence of the former, but it's also a consequence of there being only a single referee and FIFA's refusal to introduce any form of instant replay review. (Plug: my thoughts on instant replay are here.[7])

While these are familiar criticisms, I maintain 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 concerns the red card – in particular that it results in ejection of a player for the remainder of the match without allowance given for substitution.

This is an unusual complaint. But if it's a surprising charge, its connection to the issue of temporal variance might seem obscure.

Here's the connection. A central assumption undergirding the argument that basketball referees should "let 'em play" is that, 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 outcome the closer it occurs toward contest's end. Non-enforcement of the penalty at crunch time aims to rectify this imbalance.

I'm not going to suggest that soccer's red card should be brandished more reluctantly at crunch time. Unfortunately, that's not because soccer ensures that the red card exerts a constant competitive effect regardless of when issued. It's because red cards exert a greater competitive effect the earlier they are awarded. Because a red card results in ejection of the offending player and a ban on his being replaced, it entails that the offender's team play short for the remainder of the match (or until the opposition is red-carded too).

So the more time remaining at point of infraction, the greater the penalty. In effect, a red card awarded at minute 15 reads "play shorthanded for 75 minutes" whereas one awarded for the very same infraction at minute 85 reads "play shorthanded for 5 minutes." The red card thus violates the sensible principle of game

[7] Mitchell N. Berman, *Replay*, papers.ssrn.com/sol3/papers.cfm?abstract_id=1830403.

design that, presumptively, the same infraction should call forth the same penalty regardless of the time of occurrence.

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 (serious fouls, spitting, handling the ball to deny an obvious goal-scoring opportunity, etc.) were to be shorthanded for 90 minutes. In that event, the sanction would never be too high, and the fact that it would generally be too low would be unavoidable. But that's not plausible.

To be sure, what would be an optimal period of shorthandedness is extraordinarily difficult to determine. But the basic parameters are plain: Because a red card is awarded for a serious offense, the offending team should incur a significant penalty, one that meaningfully affects its prospects for victory. Yet we don't want the penalty to be virtually outcome-determinative – all the more so given the prospect (exacerbated by the prevalence of diving, by the presence of a lone referee, and by the absence of replay) that some red cards will be issued in error.

Nobody would seriously entertain a proposal to replace the penalty of ejection with the award of two goals to the opposing team. Given soccer's very low average scores and margins of victory, 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 10th minute is apt to have such a significant impact on game outcome as to contravene the competitive desideratum.

The obvious solution is for soccer 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, but the team plays shorthanded only for the remainder of that 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 re-

gardless of when the foul occurred. Many sports, not only hockey, allow a team to substitute for an ejected player after some period of penalty time. Perhaps soccer should follow their lead.

To require a team to play 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.

Figuring out what would be an appropriate period of shorthandedness would prove challenging. I'll leave that to the econometricians. I claim only that the current system that makes the competitive impact of a red card so radically dependent on its time of issuance is unlikely to dominate 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 the puzzle of temporal variance in sport.

17 Comments. //