Is There Racial Bias Among Umpires?

Phil Birnbaum

Is there widespread racial bias among umpires? In August 2007, a widely publicized academic study said the answer is yes. After taking a close look at the study, I'm not so sure.

"Strike Three: Umpires' Demand for Discrimination" is by Christopher A. Parsons, Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh. Hamermesh is the most famous of the four authors and the one quoted most often in the press reports, so I'll refer to the paper as "the Hamermesh Study." It's available for free online.[1]

Based on the results of the study, the authors (and the journalists who reported on the study) make various claims about the effects of umpire bias:

> "Specifically, an umpire will . . . call a pitch a strike about 1 percent more often if he and the pitcher are of the same race."[2]
>
> "A reasonable estimate is that a team enjoying 162 straight games of [umpire bias] advantage would win maybe one or two extra games."[3]
>
> "The data revealed that the bias benefits mostly white pitchers."[4]

I don't believe these claims are justified. A closer look at the study does show some evidence for the existence of same-race bias among umpires but does *not* show how much bias there is or where the bias lies. I believe that the quantitative conclusions quoted above are based solely on the implicit assumptions in the study, assumptions the authors may not even realize they made.

THE HAMERMESH STUDY

The authors collected pitch-by-pitch data for every regular-season game in MLB from 2004 through 2006. They classified every pitcher and every umpire into one of four groups: white, Hispanic, black, or Asian. Then, for every umpire-decided pitch (a called strike or a ball) in the sample, they noted the race of both the pitcher and the umpire. (The authors correctly write "race/ethnicity" on the grounds that Hispanic is not a "race," but I'll just say "race" to keep things simple.)

Here's the data from their table that summarizes the results. (I'm leaving out Asian pitchers because there were no corresponding Asian umpires.)

Table 1. Percentage of Pitches Called Strikes, 2004–2006

	White pitcher	Hispanic pitcher	Black pitcher	Average
White umpire	741,729	236,937	25,108	
	32.06	**31.47**	**30.61**	**31.88**
Hispanic umpire	24,592	7,323	845	
	31.91	**31.80**	**30.77**	**31.86**
Black umpire	46,825	13,882	1,765	
	31.93	**30.87**	**30.76**	**31.66**
Average	**32.05**	**31.45**	**30.62**	

The top number is the number of pitches in the sample; the bottom number is the percentage of called pitches that were strikes. It's the bottom number that's the important one here, which is why it's in bold.

If you examine the table, you'll see that there is indeed a tendency for umpires to call more strikes for pitchers of their own race. For white pitchers, they got the most strike calls when the umpire also was white. For Hispanic pitchers, they got the most strike calls when the umpire also was Hispanic. And, for black pitchers, they came within a hair of getting the most strike calls when the umpire was also black.

You'll also see from the table that white pitchers throw more strikes than Hispanic pitchers do, who in turn throw more strikes than black pitchers. Also, white umpires call more strikes than Hispanic umps do, and black umpires call the fewest strikes of all. That's not necessarily any indication of racial bias—the groups are naturally composed of different human beings, with different characteristics, and it could be just coincidence that, for instance, black umpires have smaller strike zones than white umpires. It's probably also just coincidence that the race order for pitchers just happens to be the same as the race order for umpires.

What matters is that the cells on the diagonal—the ones where the pitcher and umpire are of the same race—seem higher than they should be. For instance, white umpires called more strikes, by 0.59 percentage points, for white pitchers than for Hispanic pitchers.

But Hispanic umpires called only 0.11 percent more strikes for white pitchers than for Hispanic pitchers.

The authors ran a regression (which we'll discuss in more detail later), where they tried to predict the level of same-race bias that would best fit the nine cells of the table. They found a result of about 0.27 percentage points. That is, when facing a same race-umpire, a pitcher would be credited about one extra strike for every 400 called pitches.

However, it turned out that the result was not statistically significant. So, even though the data show more called strikes than expected when the pitcher's race matches the umpire's, the difference is so small that it could easily have happened just by chance.

MONITORING

If that were all there was, the authors of the paper would have concluded that there's no evidence of bias, and that would have been it. But they noted that there are times when umpires will find it easier to "get away" with biased calls and times when that will be harder.

For instance, in some parks, the QuesTec system electronically second-guesses umpires' ball-strike calls. For games in those parks, umpires are graded by MLB on the accuracy of their calls. In that case, you'd expect the umps' racial bias to be diminished. After all, people respond to incentives; when the umpires are punished for making the wrong call, you'd expect them to make fewer wrong calls.

Also, the authors argue, umpires can get away with more discrimination when attendance is low, because there are fewer people scrutinizing them. This I'm not sure I believe, but, as we'll see, the results do support it, so I'll go along with it.

If umpires are making the wrong calls due to racial bias, it effectively "costs more" for them to do so when they are being more heavily scrutinized. And so you'd expect them to respond to the "higher price" of discrimination by doing less of it. That's what the title of the paper is all about: umpires' "demand for discrimination" means they indulge in less discrimination when it becomes expensive to do so.

The authors found that was indeed what happened. In QuesTec parks, and games with higher attendance (when, presumably, more fans would notice), the bias disappeared—in fact, there was a bias in the *opposite* direction, as if the umpires were overcompensating. But in parks where there was no QuesTec, and where attendance was low, the apparent racial bias was much higher: statistically significant to a large degree.

Specifically, the breakdown by QuesTec status was:

+0.63	without QuesTec
-0.35	with QuesTec
+0.27	overall

That is, when QuesTec was not in effect and the umpire was of the same race as the pitcher, same-race bias resulted in a bump of 0.63 percentage points, which is one extra strike every 159 called pitches.

The breakdown by attendance was

+0.68	low attendance
-0.21	low attendance
+0.27	overall

That's one extra strike every 147 called pitches, when attendance is low and the umpire's race matches the pitcher's.

Both of the high positive results were statistically significant, which led the authors to conclude that there is indeed racial bias among major-league umpires.

REPRODUCING THE STUDY

As I said earlier, I'm unconvinced that what the authors found is really significant evidence of widespread umpire bias. Let me start by trying to reproduce the authors' results, for the low-attendance case, which showed the most evidence for bias.

For the same years the authors used (2004–6), I used Retrosheet pitch-by-pitch data to produce the low-attendance equivalent of the table 1. My results are just an approximation to what they did, but the logic that follows shouldn't be affected by the numbers being slightly different.[5]

So here's what I got for games with low attendance:

Table 2. Percentage of Pitches Called Strikes, 2004–2006 Low-Attendance Games

	White pitcher	Hispanic pitcher	Black pitcher	Average
White umpire	376,954	107,434	10,471	
	31.88	**31.27**	**31.27**	**31.73**
Hispanic umpire	10,334	2,864	258	
	31.41	**32.47**	**28.29**	**31.58**
Black umpire	23,603	6,585	695	
	31.22	**31.21**	**32.52**	**31.25**
Average	**31.83**	**31.31**	**31.28**	**31.70**

Note the similarities between this table (from my data) and table 1 (from the original study). In both cases, white umpires call the most strikes, followed by Hispanic umpires and then black umpires; in both cases,

white pitchers throw the most strikes, again followed by Hispanic and black pitchers. Also note that the numbers of pitches are in pretty much the same ratios. These factors suggest that I was able to reproduce their numbers reasonably well. (They didn't provide this particular table, which is why I had to create it.)

Also, this confirms the authors' finding that the apparent bias is higher in these low-attendance games. In table 1, white pitchers were only 0.01 above their overall average with white umpires; here, they are 0.05 above average. In table 1, Hispanic pitchers were only 0.35 above their average with Hispanic umpires; here, they are 1.16 above average. In table 1, black pitchers were 0.14 above their average with black umpires; here, they are 1.26 above average. Furthermore, in table 1, black pitchers actually got slightly fewer strike calls with black umpires than with Hispanic ones; here, however, they do significantly better with the black ump.

To make things easier to read, I'm going to redraw table 2 but without including the averages and numbers of pitches.

Table 3. Percentage of Pitches Called Strikes, 2004–2006 Low-Attendance Games (Like table 2, but with less stuff in it)

	White pitcher	Hispanic pitcher	Black pitcher
White umpire	31.88	31.27	31.27
Hispanic umpire	31.41	32.47	28.29
Black umpire	31.22	31.21	32.52

Now, clearly, and as we saw above, this table shows evidence of same-race bias. How much bias?

To answer that question, what we want is a reasonable estimate of what the table "should" look like in the absence of bias. How can we come up with that estimate?

What the authors did is to make some assumptions about how the cells get their values. Specifically, their model assumes that

> Percentage of strikes in a given cell =
> Overall percentage of strikes +
> Adjustment factor for the race of the pitcher +
> Adjustment factor for the race of the umpire +
> An extra factor for when the umpire's race matches the pitcher's race (what they call "UPM").

If you assume that the cells should all get their values that way, then you can run a linear regression to try to come up with the best estimates for all those factors, given those assumptions. It's the same technique you use to fit a straight line to a set of points—you're just trying to fit a "new table" to the "old table."

If I use the authors' technique, here's the "best fit" table I get:

Table 4. Best Fit for Table 2 Using Hamermesh Model

	White pitcher	Hispanic pitcher	Black pitcher
White umpire	31.88	31.27	31.22
Hispanic umpire	31.88	32.29	31.48
Black umpire	31.20	31.34	32.05

It's actually a pretty good fit . . . in the cells with the most pitches, the numbers hardly vary at all. The biggest differences are in the situations that didn't occur often, like black pitcher with Hispanic umpire (only 258 pitches).

The fitted matrix has roughly the same evidence of racial bias as in the original. You can see that the three diagonals still look a lot higher than they should—they're the highest numbers in their row and column, by a fair bit.

It turns out that the extra "race matching" UPM factor came out to 0.76 percentage points. (Remember that this is for my attempt to reproduce the original. It compares well to what the authors found, which was 0.68 points, again suggesting that I was reasonably able to reproduce the authors' work.) The 0.76 coefficient works out to be significant at the 5 percent level ($p = 0.0443$), which is generally the threshold for taking it seriously.

If we subtract 0.76 points from each of the diagonals (the cells where the pitcher is of the same race as the umpire), we get:

Table 5. Best Fit for Table 2 after Eliminating the Racial-Bias Estimate under the Hamermesh Model

	White pitcher	Hispanic pitcher	Black pitcher
White umpire	~~31.88~~ 31.12	31.27	31.22
Hispanic umpire	31.88	~~32.29~~ 31.53	31.48
Black umpire	31.20	31.34	~~32.05~~ 31.29

That, according to the authors' model, is the best estimate for what the results would look like if there were no racial bias among umpires. And, indeed, this updated table looks pretty unbiased. No matter who the pitcher is, Hispanic umpires call about 0.3 percent more strikes than white umpires do. No matter who the umpire is, black pitchers throw about 0.1 percent more strikes than white pitchers do. And so on.

How many pitches are affected? Well, we can multiply the 0.76 difference by the number of pitches in each of the diagonal cells. Here's what we get:

Table 6. Estimate of Number of Pitches Miscalled Due to Racial Bias under the Hamermesh Model

	White pitcher	Hispanic pitcher	Black pitcher
White umpire	+2,864		
Hispanic umpire		+22	
Black umpire			+5

By this logic, same-race umpires called 2,891 more strikes than they should have for same-race pitchers, out of 539,198 total called pitches. That suggests that about 1 pitch in 187 is affected in total. Of course, if you consider only pitches where the pitcher and umpire are of the same race, the percentage is 0.76, which is 1 in 132.

One called pitch in 132 is a bit less than one per game, I think (gotta check). Turning a single ball into a strike is worth somewhere between 0.1 and 0.14 runs. So, if we believe this analysis, having your pitcher match the umpire is worth about one tenth of a run off your ERA. That's a lot.

And you'll also notice that the advantage appears to go disproportionately to white pitchers. Even though our assumption was that all umpires are equally biased, the fact that there are so many white umpires (87) and so few minority umpires (6) means that white pitchers see an umpire who likes them almost 14 times as often as they see an umpire who doesn't like them. The study's authors conclude, therefore, that minority pitchers are disproportionately harmed by umpires' discrimination and therefore are better pitchers than their statistics suggest.

HIDDEN ASSUMPTIONS

But, as I wrote, I don't believe this is necessarily correct. There's nothing wrong with the study's math—it's the hidden assumptions that I have a problem with.

Specifically, the authors of the study insist on using the same "race bias" adjustment for each cell on the diagonal. That is, they insist on assuming that every race of umpire has exactly the same level of bias in favor of pitchers of his own race and against pitchers of other races.

Does that sound right? Not to me. I would imagine that people of different races will have different kinds and levels of bias. In almost every aspect of life affected by real, proven bigotry, it almost always goes one way. Whites used to lynch blacks; did blacks ever lynch whites? Are there gangs of gay men who roam public parks looking for handholding heterosexuals to beat up?

Even where it's obvious that two groups mutually dislike each other, does it really follow that one group will be *exactly* as biased as the other? Is a Republican boss exactly as unlikely to hire a Democrat as a Democrat boss is to hire a Republican? Even if they're equal today, what about tomorrow? When Barack Obama does something controversial overnight, don't you think that Republicans will get a lot more upset than Democrats and that the relative bias will wind up a little bit more against Democrats than it was yesterday?

If you agree that it's reasonable that the races would have different levels of bias toward each other—and even if you don't—you have to qualify the results of the study. Instead of saying

> The best estimate of racial bias is 0.76 percentage of pitches.

what you need to say is

> *If* racial bias is the same across all races, *then* the best estimate of racial bias is 0.76 percent of pitches.

Since we don't know that bias is the same across the races (and I think we have reason to believe that it's not), we can't just assume that the 0.76 percent is the right number.

And, indeed, if you relax that assumption, your conclusions can change—a *lot*, and in many different directions. There are many other ways to make the original table unbiased than by changing the three diagonals equally. Suppose we adjust it like this:

Table 7. One Way to Adjust the Results to Produce Unbiasedness

	White pitcher	Hispanic pitcher	Black pitcher
White umpire	31.88	31.27	31.27
Hispanic umpire	31.41	~~32.47~~ 30.80	~~28.29~~ 30.80
Black umpire	31.22	~~31.21~~ 30.61	~~32.52~~ 30.61

This matrix is perfectly unbiased: Pitchers get the same treatment relative to their other-race colleagues regardless of who's calling the balls and strikes. But, under this assumption, a lot fewer pitches are affected.

Table 8. Number of Pitches Affected by Bias, Based on the Adjustment in Table 7

	White pitcher	Hispanic pitcher	Black pitcher
White umpire			
Hispanic umpire		+48	-4
Black umpire		+40	+15

Here, only 106 pitches are affected—not 2,891, as in the other hypothesis. Also, instead of minority pitchers being advantaged, the exact opposite is true: Under this hypothesis, minority pitchers are the *beneficiaries* of umpire bias, not the victims!

Is there reason to believe one of these hypotheses is more plausible than the other? Maybe, but by argument, not by mathematics.

This particular hypothesis suggests that all the racial bias is shown by Hispanic and black umpires against Hispanic and black pitchers and that white umpires have no bias at all. Does that sound more likely or less likely than the Hamermesh study's hypothesis that all the races are biased equally? I don't know. But—and this is the important point—*both hypotheses are absolutely consistent with the data*.

There is literally an infinity of ways you can rejig the table to remove any evidence of bias. They all lead to different assumptions. The Hamermesh study arbitrarily chose one. There is no reason, in my opinion, to favor that one over all the others. And so, I'd argue, you can't read anything into the results.

With one exception: I believe the study does constitute evidence that there is *some* bias going on. The logic goes something like this:

Suppose that there was absolutely no bias. Then, their hypothesis, that all groups have equal bias, would be correct—all the groups would be equally biased at zero! But the study showed that, if all groups are indeed equally biased, it's unlikely to be at zero. So we have to reject the hypothesis that there's no bias going on.

But, as for the rest of the Hamermesh results . . . those are true only if bias is equal among the races. And, now that we've rejected the idea that bias is equally zero, there's no good reason to believe that bias is equal at any other level. And so there's no reason to believe that the rest of the results are consistent with what's happening in the real world.

The study has found evidence of bias but is unable to pinpoint either *how much* bias there is or *where* the bias is. Any conclusions on either of those issues are completely a result of the assumptions that went into the model.

INDIVIDUAL UMPIRES

Just as there's no justification for the assumption that all races are equally biased, there's also no justification for the assumption that all *umpires* are equally biased. That almost goes without saying. Think about the people you know in your everyday life. We all know people who are more biased than others. We know people who are a little bit biased. We all know people who believe in equal treatment. And we all know people who are so concerned about bias against race X that they argue for policies such as affirmative action. (And, depending on what kind of people you hang out with, you may even know some virulent racists.)

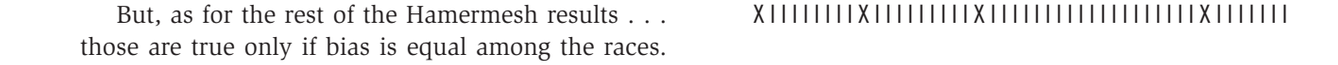
Umpires probably vary in their view of other races just as much as anyone else does. The idea that *all* umpires are biased in favor of their own race, and to exactly the same extent, doesn't seem plausible to me at all.

In that light: Is it possible that the entire effect we're seeing could be caused by only a few umpires, or even just one?

Tables 7 and 8 showed a way that the entire effect could be created by 116 miscalled pitches. Is it possible that a small number of umpires could be responsible for enough of those 116 pitches that they can push the result from statistically insignificant to statistically significant?

To check, I took every umpire in the study and compared his strike percentage with black pitchers to the strike percentage with white pitchers. If there was lots of bias, you'd expect the four black umpires to be very different from the rest—they would favor black pitchers, while the other umpires would disfavor them. If you put the umpires in order of how much they favor black pitchers, you might expect the five black umpires to be clustered at the top of the list.

They weren't *all* at the top, but they leaned toward it. Here's a graph representing where the black umpires rank in how they evaluated black pitchers:



Each vertical line represents two white umpires; each *X* represents a black umpire and a white one. As you can see, the black umpires are indeed leaning to favoring black pitchers, as there are more of them at the top (left) of the "favors black pitchers" list than the bottom. But the tendency is not huge.

Here are how the umpires rank in how much they favor Hispanic pitchers:

Again, they're closer to the top of the list than the bottom.

(Keep in mind that this doesn't necessarily mean that black and Hispanic umpires alone are biased—if white umpires are biased the other way, that would move the vertical lines toward the right side of the line, which would be enough to cause the *X*s to move left. All we can say here is that the black umpires call more strikes on black players *relative* to the other umpires—but we can't tell whether the source of the bias is the minority umps, the white umps, or a combination of both.)

Basically, these two graphs represent what the study is all about—all those numbers, charts and

regressions are just a formal mathematical way of representing what you see on these two lines. Actually, the formal method is slightly more accurate, because it takes into account the magnitude of the results, not just the rank. But, still, these *X*s and vertical lines are 90 percent of the issue.

And so, you can see that it *is* possible that one umpire could be responsible for the finding of bias. Because, as it turns out, if you remove the leftmost Hispanic umpire from the study, the leftmost *X* in the Hispanic umpire graph disappears, and the results no longer end up so significant. And if you remove the leftmost black umpire from the study, the leftmost *X* disappears from the black umpire graph, and again the results are no longer significant.

And, as you can tell just by eyeing those two graphs, if you were to remove the three leftmost minority umpires, not only would the results not be significant but the bias would be almost completely gone! The three remaining *X*s would be pretty evenly spread along the graphs.

So it's very possible that one umpire is responsible for the finding of significance and that three umpires are responsible for the entire effect.

But isn't it also possible that most, or all, umpires are still biased? Yes, of course, it's possible. It just seems unlikely that in a world where (it seems to me) there are more staunch antiracists than there are racists, a large group of umpires would all fall on the "racial bias" side. I may be wrong about this, and it's a matter of opinion . . . but, if you asked me to bet, I'd say it's much more likely to be a minority of umpires.

And, it should be said, there's a reasonable chance that there's no racial bias at all. A significance level of 0.04 isn't that extreme—it means that, one time in 25, it would happen by chance. Racial bias is a big topic in the literature, and studies that find evidence of bias are more likely to be published than studies that don't. Isn't it plausible that 25 researchers set out to find bias in baseball, and these are the only ones who did, just by chance? I think it's reasonable to argue that the jury should still be out.

CONCLUSIONS

But, anyway, there are ways to get a real answer to the question, instead of just speculating:

> Run the same study for other seasons and see if you get the same results. If you were to find the same level of bias for (say) 2000–3 that you did for 2004–6, that would be strong evidence that what the authors found is real.
>
> Look closely at the actual calls from the umpires on the left side of the line—the ones who wound up calling the most strikes for pitchers of their race. Get independent judgments about their borderline calls. If their calls look less accurate than other umpires' calls, see if that's enough to have driven the results the authors found.

Until someone actually does this further research, we can only conclude that

> There is indeed some evidence of umpire bias in favor of same-race pitchers.
>
> The bias appears at about an 0.04 level of significance.
>
> The bias appears in low-attendance and non-QuesTec situations.
>
> When attendance is higher or QuesTec is in use, the bias actually goes the other way.
>
> We can't tell which umpires are biased, how many umpires are biased, or even what races of umpires are biased.
>
> We can't tell how many pitches are affected by the bias. It could be as few as 116, or it could number in the thousands.
>
> We can't tell which races of pitchers are beneficiaries of the bias and which races are harmed by it.

So: there's statistically significant evidence for bias, but we don't know which races are affected, how many umpires have it, or how strong the bias is. Quite unsatisfying for fans, reporters, and researchers alike—but I think that's all this study gives us. ■

Notes

1. Christopher A. Parsons, Johan Sulaeman, Michael C. Yates, and Daniel S. Hamermesh, "Strike Three: Umpires' Demand for Discrimination," at Social Science Research Network, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1318858.
2. Alan Schwarz, "Keeping Score; A Finding of Umpire Bias Is Small but Still Striking," *New York Times*, 19 August 2007.
3. Ibid.
4. Katie Rooney, "Are Baseball Umpires Racist?" *Time*, 13 August 2007.
5. Technical note: To save time, I only approximated what the Hamermesh authors did. I didn't correct for count, score, home/road pitcher, batter, or umpire, which they did. I selected games by actual attendance (less than 30,000) instead of the study's 70 percent of capacity. For umpires, I considered only Angel Hernandez and Alfonso Marquez as Hispanic, and C. B. Bucknor, Laz Diaz, Chuck Meriwether, and Kerwin Danley as black. The original study included one additional Hispanic umpire and one additional black umpire, but I don't know which ones those are. Also, for Hispanic pitchers, I used only those born in one of the countries listed in the study; and, for black pitchers, I used only those in the list "African-Americans in MLB, 2007" at Black Voices (www.blackvoices.com). But I got similar results both the in relative number of pitches seen, and in the effect of those pitches. So I figure it's close enough.