Draft 3.0 1.18.25 [subject to revision]

Fielding Before and After Baseball's Great Transformation

Xavier Fünf[*]

Abstract

This paper presents evidence of a dramatic decrease in the importance of team fielding quality in major league baseball. Over the course of the last three decades, the share of variance in runs allowed that is explained by fielding has steadily declined as the share accounted for by fielding-independent pitching has steadily risen. The paper uses a variety of non-digital and digital fielding metrics, including MLB's Statcast, to chart this trend. It also illustrates the practical effect of it on season-long outcomes and on the evaluation of individual player WAR.

Introduction

Baseball has changed. Once characterized by the skillful dance of pitchers and fielders seeking to interrupt the fine-tuned consistency of keen-eyed batsmen, major league games today feature a succession of violent showdowns between lab-grown strikeout assassins and swing-happy home run bombardiers (McCullough, 2024; Verducci, 2017) (Fig. 1).

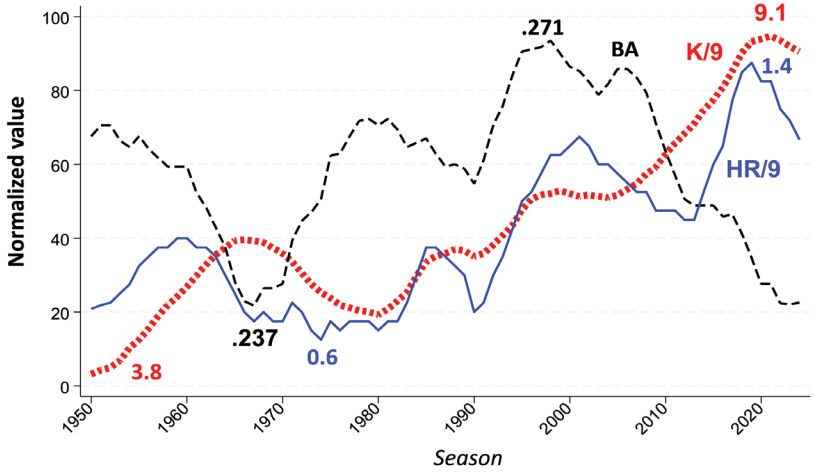


Fig. 1. MLB Trends in batting averages, strikeouts per 9 innings pitched, and team home runs per game. Values normalized on 0-100 scale with high and low points marked for reference. Trend lines reflect 5-year rolling averages.

My principal goal in this article is to examine a consequence of this "Great Transformation," one that flows naturally from it but that has nonetheless evaded systematic measurement: the diminishing importance of fielding in major league baseball. In a game in which home runs and strike outs matter more and more, we should expect differences in the proficiency of teams' fielders to matter less and less. I will

[*] Research associate, bbcardstats.com. Direct correspondence to xavier_funf@bbcardstats.com. I am grateful to Charles Pavitt for comments on an earlier draft.

present evidence to substantiate this "fielding shrinkage" hypothesis and to quantify the practical significance of fielding's diminished role.[1]

The paper also has a secondary aim: to empirically test pre-digital fielding-quality metrics. Indeed, the success of my primary goal presupposes the validity of these measures. The evolution of the game toward the dominance of strikeouts and home runs has occurred largely (but not entirely) since the advent of fielding analytics driven by video data of the location, speed, and trajectory of batted balls. Substantiating the shrinkage hypothesis, then, requires determining the relative impact of fielding and pitching in earlier periods of baseball history.

In the course of this analysis, I will also report some collateral findings. They relate principally to performances measures that currently figure in calculation of player WAR.

Methods

To conduct my examination, I assembled data on the impact of fielding and pitching across every season of the American and National Leagues since 1900. The basic strategy was to examine the incremental contribution of fielding to runs allowed at the team level after taking account of team pitching.

To assesses the latter, I used FIP or fielding-independent pitching (Tango, 2004; McCracken, 2001), Because it measures pitching quality independent of fielding, FIP is ideally suited as a foundation or control variable for assessing the incremental contribution of fielding to runs allowed. In addition, FIP features strikeouts and home runs given up, and can thus be used to gauge the impact of the two game outcomes one would expect to be most responsible for any decline in the impact of differences in team fielding proficiency.

For fielding metrics, I used two pre-digital measures and three digital ones. The pre-digital systems were the fielding-runs allowed components of Total Zone Rating (TZR) and Defensive Efficiency Record (DER). Developed by Sean Smith (2024) via his painstaking coding of Retrosheet data, both of these measures rely principally on the proportion of balls in play that a proximate fielder converts into outs. TZR is relatively more discriminating, excluding from its tallies outs that fielders can be expected to make more routinely. The TZR fielding-runs measure is used to compute player WAR by both Baseball Reference and FanGraphs before 2003. I used Baseball Reference's "rfield" measure as my source for TZR data.

I supplemented the analysis with DER, which is available from Smith's Baseballprojection.com site, for two reasons. First, as Smith (2024) has recounted, the TZR ratings for the decade of the 1990s were attenuated by his temporary substitution of data from "Project Scoresheet" for his own Retrosheet coding. Second, whereas TZR is not available for seasons after 2003, Smith continues to update DER. The latter's fielding-runs saved component can thus be used to establish directly how well pre-digital fielding measures correspond with the digital measures that cover seasons since then.

From the 2003 season onward, both Baseball Reference and FanGraphs use digitally derived fielding metrics to calculate player WAR. Baseball Reference uses Defensive Runs Saved (DRS). For the seasons of 2003 to 2015, FanGraphs employs an adjusted version of Ultimate Zone Rating (UZR) (Lichtman, 2017); for 2016 to 2024, it uses the runs-prevented element of Statcast's Outs Above Average scheme. Both DRS and UZR use data generated by Baseball Info Solutions, while Statcast uses data generated by Major League Baseball itself. DRS, UZR, and OAA all derive runs saved from models of the

[1] I am indebted to Sean Smith for putting me onto this hypothesis in a response to preliminary analyses of the data presented in this paper.

probability that a ball hit to a particular sector of the field will be turned into an out, although Statcast uses more fine-grained information on the speed and trajectory of the ball (Lichtman, 2017; Schoenfeld, 2016). For my analyses, I used the DRS data available from Baseball Info Solution's Fielding Bible website, the UZR data available on FanGraphs, and the OAA data available from MLB's Baseball Savant website.

All the reviewed measures are conceptually and functionally similar. They use diverse sources of evidence. But all translate rates of successful fielding of balls in play into estimates of runs saved. They are thus well suited for comparative analysis.[2]

The basic analytical strategy involved regressing team runs allowed, first, on FIP alone, and then on FIP together with each of the five fielding measures, one at a time. This method makes it possible to identify the incremental contribution of each fielding measure to the percentage of variance explained (R^2) in runs allowed.

These analyses use team-level versions of the fielding measures, formed by aggregating teams' individual players' runs-saved scores. The explained variance attributable to team-level fielding measures so formed depends on the validity and accuracy of the individual-level scores summed to form them. Accordingly, the incremental R^2s of the models necessarily reflect the explanatory power of the relevant fielding measures at the individual level. Indeed, because *actual* fielding runs saved at the individual level are not observable, the power of aggregated individual scores to explain team-level variance in runs allowed is the only method available for empirically validating measures of individual fielding proficiency.

The performance of pre-digital measures

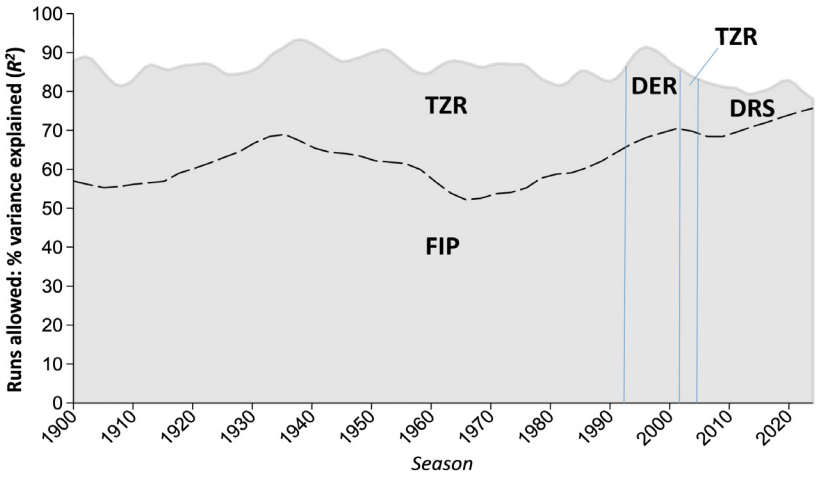


Fig. 2. Relative contributions of pitching and fielding to variance in team runs allowed, 1900-2024. AL/NL only. Gray region reflects overall variance explained (R^2) for single season models in which runs allowed are regressed on FIP and the indicated fielding measure; the area below the dashed line represents the contribution of the former, the area above the latter (Stat.App. Note 1). DER is substituted for TZR for 1990-1999 due to the latter's reliance on inferior data in that period (Smith, 2024). DRS is used for 2003 to 2024 as the digital metric with highest R^2 (Stat.App. Table 2). Local polynomial smoothing applied to changes in yearly R^2s for graphic purposes.

As reflected in Figure 2, from 1900 to 1990, TZR consistently explained over 30% of the variance in team runs allowed after controlling for FIP. Together, FIP and TZR consistently explained 80% to

[2] Preliminary analyses were also performed on one additional pre-digital measure, Defensive Regression Analysis (Humphreys 2011). See Stat.App. Note 4.

85% of the variance (consult the on-line Supplemental Information ["Stat.App."] for more fine-grained details).

Derived straightforwardly from individual fielding performance records, there is nothing except differences in fielding skill that TZR could be understood to be adding to the power of the models to explain differences in runs allowed. The impressively high degree of variance explained in these models furnish compelling evidence of the validity and power of TZR as a measure of fielding proficiency in the twentieth century.

3.2. The ascendance of FIP, the decline of fielding

Although the trend begins as early as 1980, the expanding impact of FIP on differences in team runs allowed accelerates through the 1990s and has continued to grow every decade thereafter. From a value of around 50% to 55% of variance explained through the 1960s and 1970s, FIP balloons to 77% for the last decade (Stat.App. Table 1; Fig. 2). Concurrently, the incremental contribution of fielding runs saved becomes progressively smaller. This trend strongly supports the principal "fielding shrinkage" hypothesis.

How much has the impact of differences in teams' fielding skills contracted? The answer varies depending on what runs-saved measure one employs.

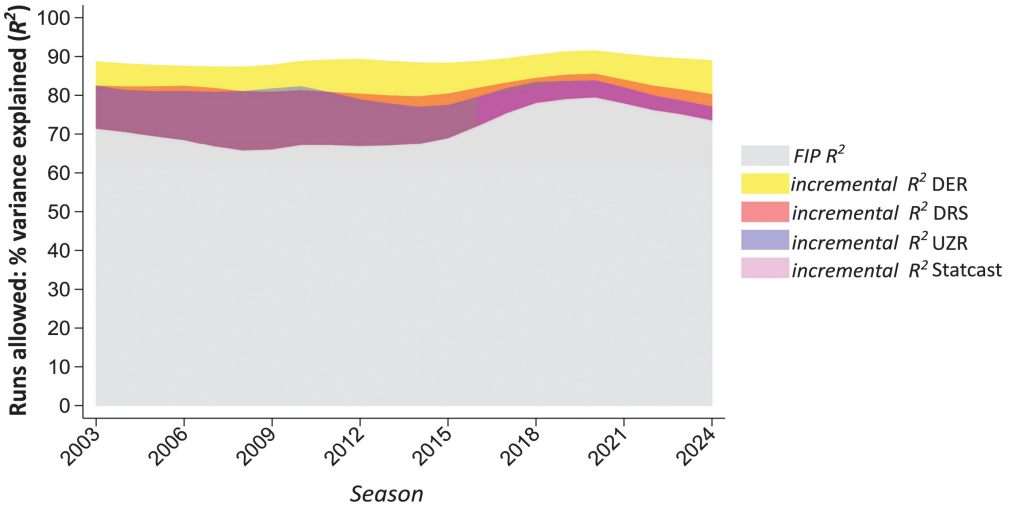


Fig. 3. Relative incremental contributions of fielding measures to variance in runs allowed, 2003-2024. Gray region reflects overall variance explained (R^2) for single season models in which Runs Allowed are regressed on FIP alone; the overlapping colored transparency regions reflect the incremental contributions when DER, UZR, DRS, and Statcast (post-2015). Local polynomial smoothing applied to changes in yearly R^2s for graphic purposes (Stat.App. Note 1).

As indicated, digital fielding-proficiency measures are now used to determine player WARs. From 2015 to 2024, DRS accounted for 7% of the variance in runs allowed; UZR and Statcast (the latter from 2016) each accounted for about 4% (Stat.App. Table 1). Over the course of the 1960s and 1970s, fielding accounted for 30% of the difference in team runs allowed (Stat.App. Note 1). Accordingly, it could be estimated that differences in fielding have diminished over 75% since then as a contributor to differences in the number of runs that teams allow.

While still substantial, the impact is smaller if we assess the current effect of team fielding differences with Smith's DER measure. Over the last decade, DER explains 13% of the variance in team runs

allowed (Stat.App. Table 1). Using this estimate, then, we might conclude that fielding differences have tailed off around 55% since FIP began to climb in significance.

DER's explanatory advantage over digital measures extends over the entire period in which the latter have been in use. Over the seasons from 2003 to 2024 combined, DER accounts for 17% of the variance in team runs allowed, as compared with 10% for DRS and 8% for UZR (Stat.App. Table 1). These results fortify confidence in the inferences that can be drawn about the relative importance of fielding over the period that spans the use of pre-digital and digital measures. Indeed, the power of DER to ferret out these effects suggests that it would be a mistake to discount pre-digital measures as a source insight not only for the period before the advent of digital ones but for the seasons that have occurred thereafter as well.

3.2. Gauging the effect of fielding shrinkage

Universally attested to by all the examined measures, the phenomenon of fielding shrinkage does not mean that it is now irrelevant who occupies teams' defensive positions. Obviously, if a team selected its fielders at random from a stadium of fans, it would be demolished. But so long as they confine themselves to players of reasonable defensive competence, teams whose fielders lag behind others in quality are unlikely to suffer much.

As a practical illustration of this point, we can use the analyses conducted here to re-examine the outcomes of various historical pennant races (Table 1). In 1948, for example, the Cleveland Indians and Boston Red Sox finished deadlocked with identical 96-58 records, resulting in a one-game playoff won by the Indians (Epplin, 2021). But had fielding borne the same significance that season that it did in 2024, it's likely no tiebreaker would have been necessary. Fielding shrinkage would have shaved *44* runs from the runs-saved advantage that shortstop Lou Boudreau (+20 TZR), second baseman Joe Gordon (+16), third baseman Ken Keltner (+12), and center fielder Larry Doby (+11) had secured the Indians in 1948. Based on the oft-confirmed finding that 10 runs scored or avoided over a season equate to approximately 1 win (e.g., Thorn and Palmer, 2015), the Red Sox would have been expected to win the pennant by 4½ games.

	Pitching/Fielding Run Advantage		
	then	now	**Total Run Δ**
1948 Indians vs. Red Sox	95	51	**-44**
1962 Giants vs. Dodgers	-20	-52	**-32**
1967 Red Sox vs. Twins	-18	-60	**-42**

Table 1. "What if" replays. "Then" and "now" pitching and fielding run advantages for pennant-winning team reflect regression-driven Monte Carlo simulations of the impact of teams' respective FIP and "rfield" (TZR/DRS) scores. Simulations based on proximate four-season pooled data (Stat.App. Note 2).

The 1962 Giants likewise won the pennant in a post-season playoff, defeating the Dodgers 3 games to 2 (Krell 2021). The Giants would likely not have had the chance to advance, however, but for their superior defense. Led by Willie Mays's stellar performance in center field (+20 TZR), the Giants' 49 fielding-runs-saved advantage would have largely neutralized the Dodgers formidable pitching advantage. Played in 2024, however, the Dodgers superior pitching would have swamped the Giant's glove men, tipping the runs-allowed balance in the Dodgers favor by 52 runs. The resulting net 32-run cushion should have been enough to eliminate the Giants from contention.

In 1967, three teams—the Red Sox, Twins, and Tigers—entered the last weekend of the season with realistic shots to win. When the dust settled, the Red Sox emerged victorious, a game ahead of both the Tigers and the Twins (Bright 2018). Again, fielding was arguably decisive, at least as between Boston and Minnesota. Owing principally to the sure-handedness of left fielder Carl Yastrzemski (+23 TZR) and the reliable glove of shortstop Rico Petrocelli (+8), the Red Sox boasted an expected 67 runs-saved fielding advantage over the Twins, who had to endure the sloppy ball handling of third baseman Rich Rollins (-10), outfielder Bobby Allison (-8), and first baseman Harmon Killebrew (-5). The Twins, however, enjoyed the pitching edge: a 2.97 FIP versus the Red Sox's 3.50. The combined excellence of Jim Kaat (who recorded a league-best 2.55 FIP) and the solid work of Jim Perry (FIP 3.11) made the Twins staff harder to hit, notwithstanding Jim Lonborg's Cy Young campaign. In a 2024 replay, the Twins pitching advantage would far outpace the Sox's fielding lead—to the tune of a net gain of 40 additional runs saved, and thus an expected gain of approximately 4 wins (Table 1). The Twins would in all likelihood have returned to the World Series for the second time in three years.

These "what if" analyses are obviously confounded by differences in how these teams would have adjusted their rosters and lineups for a 2024 game environment. But that is exactly what an empirically enriched thought experiment of this sort tells us—how much teams *should* now discount the value of superior fielding relative to its impact in the middle of the last century.

Runs-saved calibration

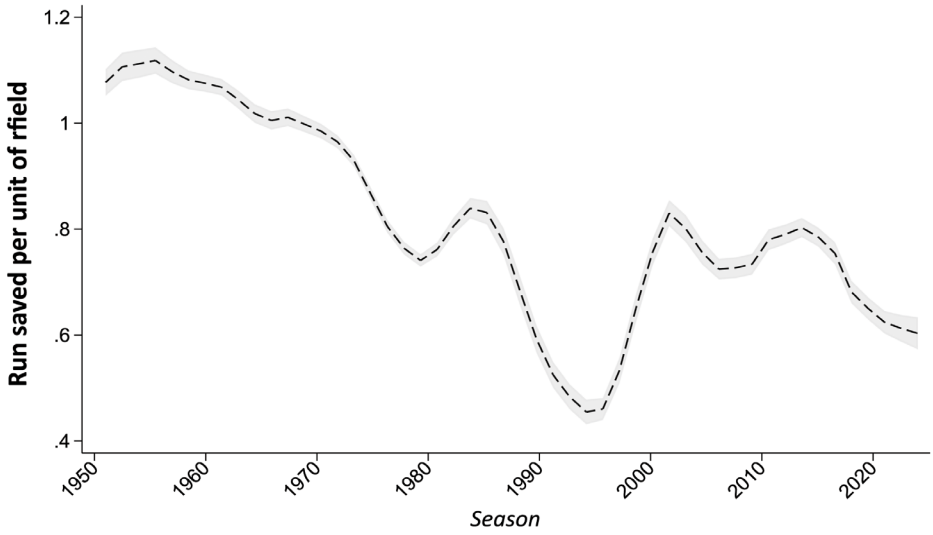


Fig. 4. Calibration of Baseball Reference runs-saved measures. Plotted line reflect the correspondence between 1-unit of the rfield runs-saved measure and actual runs saved. Y-axis value are derived from single-season regression models of runs allowed on FIP and rfield (Stat.App. Note 1). Polynomial smoothing used for presentation. Note that the performance of rfield in the 1990s is not a consequence of miscalibration but rather the impact of the temporary use of inferior data in the calculation of TZR, as recounted by Smith (2024).

A fielding "runs saved" measure is meant to convey what it says: a count of the runs averted by virtue of the quality of a fielder. My investigation, however, revealed that the actual runs prevented by a unit of fielding runs saved has tended to fall short of that, a trend that has grown over time. For example, whereas a unit of Baseball Reference's "rfield" was worth about 1 run in the 1960s (when based on TZR), it was worth only 0.55 on average in the decade ending in 2024 (Fig. 4).

This dynamic—which I'll call "runs-saved inflation"—is only an indirect consequence of fielding shrinkage. It is true that fielding matters less now than it did for most of the last century, before the contest between strikeouts and home runs came to dominate the game. But this development could easily be accounted for by appropriately crediting fielding skill with what it is actually worth when tallying team runs allowed. "Runs saved inflation" reflects the failure to recalibrate fielding-runs measures as differences in pithing quality have assumed a progressively larger role in determining differences in the number of runs that teams surrender.

Indeed, this calibration problem is less severe with other measures. Over the last decade, a unit difference in the DER runs-saved measure corresponded to an average of 0.79 runs. A unit of UZR was worth 0.89 on average over that period (Fig. 5).

A one-run increment in Statcast's measure of fielding proficiency was worth 0.96 over the period from 2016 to 2024 considered as a whole. The trajectory of the measure, however, is characterized by a sharply inverted "V": an initial overvaluation of runs saved, followed by a steep decline (Fig. 5). That pattern extends "runs saved inflation" to "DEF," the Statcast-based fielding-runs measure that FanGraphs uses to calculate player WAR: over the last two seasons, a unit of DEF is worth an average of 0.66 actual runs (Stat.App. Note 1).

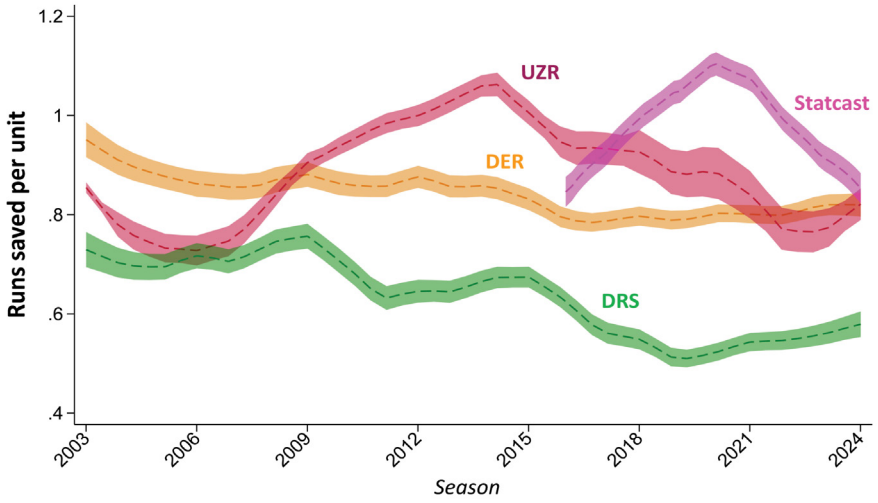


Fig. 5. Calibration of UZR, DRS, DER, and Statcast runs-saved measures. Plotted lines reflect the correspondence between 1-unit of indicated the runs-saved measure and actual runs saved. Y-axis values are derived from single-season regression models of runs allowed on FIP and on each of the indicated runs-saved measure (Stat.App. Note 1). Polynomial smoothing used for presentation.

Because "runs-saved inflation" bears only on the units in which fielding proficiency is measured and not on the validity of the systems being used to measure it, it does not pose any genuine barrier to empirical assessment of baseball performance. But to the extent that the runs-saved measures used to calculate player WAR are treated at face value, they risk overstating both the credit due good fielding, and the blame due bad, in determining teams' fates. In particular, "runs saved inflation" tends to distort historical comparisons of fielding performance.

This point can be practically illustrated by considering the career fielding-runs-saved ranking of third basemen (Table 2). Brooks Robinson occupies the number one spot on Baseball Reference's runs-saved list; he not only played before "runs saved inflation" started to degrade runs-saved scores but his rfield score is so far ahead of everyone else that no amount of adjustment could possibly dislodge him from his place atop of the list. But three of the next four places are assigned to third basemen who played

most or all of their careers in the current century: Adrian Beltré, Scott Rolen, and the still active Nolan Arenado. Because "runs saved inflation" was raging by that point, they are players whose contributions to avoiding runs is most likely overstated.

With an eye toward correcting this possible distortion, I re-examined the Baseball Reference rfield scores of these players as well as those of Buddy Bell and Clete Boyer, players whose rfield rankings—fourth and sixth, respectively—are most likely to have been adversely affected by "runs saved inflation." My principal adjustment was to replace the post-2000 rfield scores for Beltré, Rolen, and Arenado with the *actual* runs-saved associated with their scores, as determined by regression analyses for the relevant seasons. Because, as discussed, the TZR scores used to compute rfield scores for the 1990s are not reliable (Smith, 2024), I excluded those seasons from the analysis and adjusted upward Beltré and Scott Rolen runs-saved totals by an amount that reflects the (small) proportion of career games they played, respectively, before the 2000 season (Stat.App. Note 1).

The results are reported in Table 2. As can be seen, Clete Boyer is elevated to number two on the all-time fielding-runs saved list. Boyer leapfrogs Beltré, Rolen, and Arenado. He also edges ahead of Bell, both as a result of a modest undervaluation of fielding impact by TZR for part of the 1960s and a modest degree of overvaluation during parts of Bell's career. Nevertheless, Bell also overtakes Rolen, securing third on the revised all-time runs-saved list.

	Beltré	**Rolen**	**Bell**	**Arenado**	**Boyer**
Baseball Reference rfield	202*	147*	174	162	161
calibration adjustment	-52	-31	-37	-42	10
adjusted total	163†	151†	137	120	171
Unadjusted ranking	2	3	4	5	6
Adjusted ranking	**4**	**5**	**3**	**6**	**2**

*Score excludes seasons played in 1990s.
†Total adjusted upward proportional to games played in 1990s.

Table 2. Adjusted runs saved. The "adjusted total" and "adjusted ranking" are based on removal of estimated impact of "runs saved inflation" (Stat.App. Note 1).

For these analyses, I used the TZR and DRS runs-saved measures that informed Baseball Reference's rfield measure. But as a check, I compared these results to ones derived from the DER measure. In addition to explaining more variance than DRS for the period after 2002, DER's run-saved measure displays a more faithful relationship to actual runs-saved: over the last decade, a unit difference in the DER runs-saved measure consistently corresponds to an average of 0.79 runs (Fig. 5). In accord with my adjusted analysis, Basebeallprojection.com's DER-based system ranks Boyer second in third-base fielding runs saved (166), ahead of Beltré, Rolen, Arenado, and Buddy Bell.

Conclusion

Baseball has always been and remains a beautiful game. But at least one dimension of its beauty no longer has the significance that it once had. For most of the twentieth century, superb fielding—such as that displayed by Orioles of late 1960s and early 1970s—was appropriately recognized as a key element of team success over the course of a 154- or 162-game campaign. Exquisite glove work might still garner admiration. But the truth of the matter is, it just doesn't matter nearly so much in today's game environment, in which strike outs and home runs have assumed a preeminent place. Fielding accolades are effectively now awarded on basis of style points that satisfy aesthetic sensibilities that are growing progressively more remote from game outcomes.

This development might be considered unfortunate, but it is almost certainly too late in the day to change it. What can be improved upon, however, are the quality and the calibration of measures of fielding, the imprecision of which have partially obscured this element of the great transformation that has characterized major league play in the twenty-first century.

There is also room for improvement of digital fielding measures generally. The point of this paper was not to determine which fielding measure is "best," but only to assess how the escalating significance of strikeout-pitching versus home-run hitting has affected the consequence of differences in team fielding quality. But in testing the power of pre-digital measures to support such assessments, it was revealed that measures like TZR and DER remain more reliably attended to the impact of fielding play today than do the formulas that drive digital fielding metrics. There is nothing necessary about that. On the contrary, with additional empirical fine-tuning of the models used to process the fine-grained evidence generated by digital technologies, the latter seem destined to emerge as superior (Lichtman, 2017). The analysis here, I hope, will contribute to the project to refine these metrics.

References

Basco, D. & Zimmerman, J. Measuring defense: entering the zones of fielding statistics. Baseball Research Journal, 39:83-98 (2010).

Bright C (2018) The 1967 American League Pennant Race: Four Teams, Six Weeks, One Winner. McFarland, Incorporated, Publishers..

Epplin L (2021) Our Team: The Epic Story of Four Men and the World Series That Changed Baseball. Flatiron Books.

Humphreys MA (2011) Wizardry : baseball's all-time greatest fielders revealed. Oxford ; New York: Oxford University Press.

Krell D (2021) 1962: Baseball and America in the Time of JFK. Nebraska.

Lichtman M (2017) Defensive evaluation. Handbook of Statistical Methods and Analyses in Sports. Chapman and Hall/CRC, pp.83-104.

McCracken V (2001) Pitching and Defense: How much Control Do Hurlers Have. Baseball Prospectus, Jan. 3, 2001. Available at: https://web.archive.org/web/20060613054445/http://baseballprospectus.com/article.php?articleid=878.

Andy McCullough (2024) Missing Bats: How an obsession with strikeouts upended the balance of baseball. Athletic, June 24, 2024. Available at: https://www.nytimes.com/athletic/5537852/2024/06/24/mlb-missing-bats-baseball-strikeout-pitch-tracking/.

Schoenfeld B (2016) Can New Technology Bring Baseball's Data Revolution to Fielding? New York Times Magazine, Oct. 2, 2016.

Smith S (2024) War in Pieces. North Haven, Conn.: Independent.

Tango M (2004) Defensive Responsibility Spectrum (DRS). Available at: https://web.archive.org/web/20040804215058/http://www.tangotiger.net/drspectrum.html.

Thorn J, Palmer P and Law K (2015) The Hidden Game of Baseball. University of Chicago Press.

Verducci T (2017) Baseball's pressing question: What happens to a sport when nothing happens? Sports Illustrated, June 26, 2017.

Statistical Appendix

Note 1. Individual-season regression models

The principal statistical analyses for the paper consist of a set of multivariate regressions, some of which cover all and others subsets of every National League and American League season from 1900 to 2024. Data from both leagues were combined based on the judgment that inter-league differences are unlikely to be of sufficient practical consequence or theoretical interest to sacrifice the loss of statistical power associated with splitting the samples for the each season's models in half.

The models consist of two steps: first the regression of team runs allowed on FIP, and second, the regression of team runs on FIP and one of the fielding measures analyzed in the paper. That generates five distinct sets of season-by-season models of this form, one corresponding to each of the following fielding metrics: (1) rfield, the Baseball Reference runs-saved measure, which uses TZR for 2000-2002 and a modified version of DRS for 2003-2024; (2) the DER total runs saved measure, which was obtained at Sean Smith's Baseballprojection.com site; (3) Fielding Bible's raw DRS measure, which covers the seasons 2003-2024; (4) raw UZR runs-saved, obtained from FanGraphs, which covers seasons from 2003 to 2024; and (5) the Statcast total runs saved measure (2016-2024), which was obtained from Baseball Savant. Similar data were also collected (6) for FanGraph's DEF fielding measure, which largely tracks the measures it is based on (TZR through 2002, UZR from 2003 to 2015, and Statcast from 2016 to 2024).

Because it would be infeasible to reproduce and nearly impossible to comprehend the model outputs in a conventional table, they are instead reported in a downloadable excel file. The file has six separate workbooks, one for the model associated with each of the five fielding measures featured in the paper and another for FanGraph's DEF measure. Each workbook reports in separate columns: (a) the beta weight for the FIP-alone model (b_FAM); (b) the t-statistic associated with that model's predictor beta weight (t_b_FAM); (c) that model's constant (cons_FAM); (d) the *t*-statistic associated with that constant (t_cons_FAM); (e) the R^2 associated with that model (*R*2_FAM); (f) the beta weight for FIP in the model that includes FIP and the indicated fielding measure (b1_FPFM); (g) the *t*-statistic associated with that beta (t1_FPFM); (h) the beta weight for the fielding measure (b2_FPFM); (i) the *t*-statistic for that beta (t2_FPFM); (j) the constant for that model (cons_FPFM); (k) the *t*-statistic for that constant (t_cons_FPFM); (l) the R^2 for that model (R2_FPFM); and finally (m) the incremental R^2 associated with the addition of the fielding measure (R2i_FM). Each row of a given worksheet includes this information for the specific model fit to the indicated major league season.

These data were the basis of the following results and findings described in the paper:

(A) Figure 2 is based on the models for 1900-1989 and 2000-2002 in the "rfield" sheet; on 1990-2000 in "DER" sheet; and on 2003-2024 in the "DRS" sheet.

(B) Figure 3 is based on the models for 2003-2024 in the "DER" sheet; on 2003-2024 in the "UZR" sheet; 2003-2024 in the "DRS" sheet; and 2016-2024 in the "Statcast" sheet.

(C) The actual-runs-allowed results reported in connection with the discussion of runs-saved calibration and in Figures 4 and 5 reflect estimates based on the FIP-plus-fielding-measure model for the indicated fielding metric.

(D) In Table 2, the paper presents estimates of the impact of "runs saved inflation" on Baseball Reference's all-time runs saved list for third basemen. The estimates were formed principally by substituting for the season-by-season Baseball Reference rfield scores of the featured third base-

men the "actual runs allowed" estimates generated by the FPFM models ("rfield" sheet) for relevant seasons. In addition, as indicated in the text, scores for the 1990s were excluded and the final adjusted scores increased proportional to the number of career games the indicated third baseman played in the 1990s. For Adrian Beltré, the proportional upward-adjustment factor was 7.7% (229 of 2993 games); for Scott Rolen, the amount was 22.8% (465 of 2038 games).

Note 2. Multi-season regression models

The paper reports the relative explanatory power of DER, DRS, UZR, and Statcast for the seasons spanning 2003 to 2024. The reported results reflect the regression model outcomes in Stat.App. Table 1.

	Seasons								
	2003-2024				*2015-2024*				
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(5)
z_FIP	**0.85**	**0.73**	**0.77**	**0.82**	**0.88**	**0.75**	**0.78**	**0.86**	**0.84**
	(40.97)	(53.45)	(45.05)	(46.95)	(30.24)	(38.92)	(38.92)	(34.27)	(30.54)
z_DER		**-0.43**				**-0.38**			
		(-31.42)				(-19.61)			
z_DRS			**-0.32**				**-0.28**		
			(-18.85)				(-10.99)		
z_UZR				**-0.28**				**-0.21**	
				(-16.16)				(-8.34)	
z_Statcast*									**-0.20**
									(-7.18)
Cons	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
R^2	0.72	0.89	0.82	0.80	0.77	0.91	0.84	0.82	0.81
ΔR^2		**0.17**	**0.10**	**0.08**		**0.13**	**0.07**	**0.04**	**0.04**

Stat.App. Table 1. Regression models, 2003-2024 & 2015-2024 seasons. *Statcast model covers seasons from 2016 to 2024 only. Outcome variable is runs allowed per game. Variables are standardized by season to confine model to estimates of the impact of predictors on variance in runs allowed within seasons as opposed to variance across them and to remove the effect of inter-season variability unrelated to the impact of the predictors on the outcome variable (Schell 1999, 2005). Beta *t*-statistics are in parentheses. Bolded predictors and ΔR^2s are significant at $p < 0.01$.

Whereas the single-season regression models use raw or untransformed values for the output and predictor variables, the multi-season regressions here use season-standardized values of the same. This transformation puts the predictors and explanatory variables for all seasons on a common scale, thereby removing bias and noise associated with inter-season influences that affect runs scored independently of the relationship of the predictors (Schell 1999, 2005).

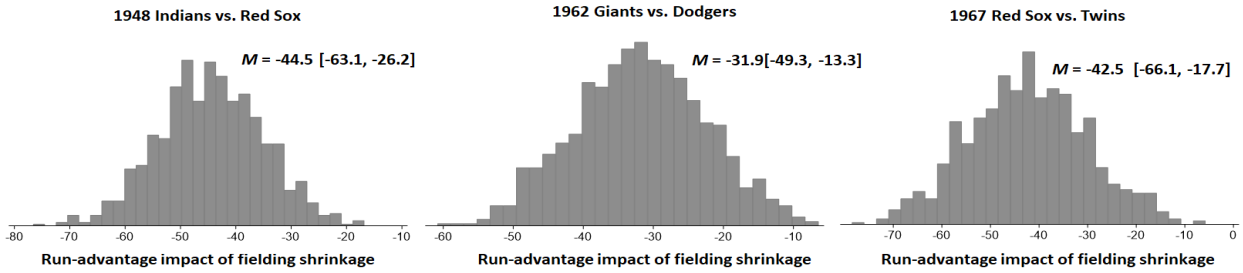
Note 3. "What if" and Table 1

The paper illustrates the impact of "fielding shrinkage" with a series of estimates of how the shifting importance of fielding and pitching would have affected the runs-allowed differentials for teams involved in three past-season pennant races. To assure the analyses were not unduly influenced by random variation associated with the particular seasons being compared, the "then" estimates were based on pooled data across the four seasons concluding in the past season in question, and the "now" estimates on pooled data across the 2021 to 2024 seasons. Baseball Reference rfield was chosen as the fielding measure because it reflects the highest R^2 pre-digital measure for the historical seasons (TZR), and the highest R^2 post-digital one (DRS) for the contemporary-period seasons. Regression models were first fit to the

"then" and "now" periods (Stat.App. Table 2). The models were then used to drive the Monte Carlo Simulations using the teams' respective FIP and rfield scores (Stat.App. Fig. 1).

	Seasons			
	2020-24	1945-48	1959-62	1964-67
FIP	**183.27**	**156.51**	**158.43**	**160.25**
	(21.36)	(13.97)	(12.47)	(14.29)
Rfield	**-0.57**	**-1.36**	**-1.16**	**-1.03**
	(-5.47)	(-10.92)	(-10.18)	(-11.27)
Cons	-41.03	48.54	90.65	86.24
	(-1.14)	(1.10)	(1.80)	(2.22)
R^2	0.82	0.86	0.85	0.86

Stat.App. Table 2. Regression models, "What if" analyses. Outcome variable is runs allowed. Beta t-statistics are in parentheses. Bolded predictors and ΔR^2s are significant at $p < 0.01$.



Stat.App. Fig. 1. "What if" Monte Carlo simulation outputs. Regression-model parameters (Stat.App. Table 2) drove a Monte Carlo simulation using the indicated teams' team FIP and rfield scores. One thousand simulations of the differences between "then" and "now values were run for each "replayed" pennant race. Mean run-differentials are reported with values at 2.5 and 97.5 percentiles indicated in brackets.

Note 4. Defensive Regression Analysis

Defensive Regression Analysis (DRA) (Humphreys 2011) is another non-digital system for estimating fielding runs saved. DRA uses a sequentially linked set of regression analyses to determine the rate at which different types of balls hit in play are turned into outs by specified fielders as well as the runs-prevented consequence of the resulting rates.

Neither Humphreys (2011) nor any commercial firm or individual researcher has made DRA scores publicly available for AL and NL teams. Nevertheless, Baseball Reference uses DRA to compute the fielding-runs saved scores (rfield) associated with its Negro League WAR calculations (Baseball Reference undated). I assembled these data and conducted an analysis of it akin to the ones reported in this paper.

Three sets of analyses are presented. The first involves Negro National League I. In operation from 1920-1931, NNL I generally fielded 6-8 clubs, which played from 50-80 games per team. Negro National League II (1933-1948) was the longest continuously running league. It usually fielded half a dozen teams, which tended to play around 50 games per season (Heaphy 2003). Analyses of the DRA fielding measure was performed for this league as well. Considerably less data are available for the remaining

leagues (the American Negro League, the Eastern Colored League, the Negro American League, and the East-West League). Accordingly, to assure adequate statistical power, data from these leagues were combined.

The analyses are reported in Stat.App. Table 3. Ranging from 0.11 to 0.14, the incremental R^2s for the DRA-based fielding runs saved analysis, confirms the validity of DRA. Nevertheless, the contribution DRA makes to the variance in team runs explained is substantially lower than that made by TZR and DER over the AL and NL seasons played before the advent of digital measures.

	NNL I		*NNL II*		*other leagues*	
	(1)	(2)	(3)	(4)	(5)	(6)
z_FIP	**0.72**	**0.55**	**0.40**	**0.29**	**0.26**	0.19
	(6.90)	(5.17)	(3.23)	(2.47)	(2.10)	(1.54)
z_rfield		**-0.37**		**-0.39**		**-0.37**
		(-3.47)		(-3.35)		(-3.07)
N	46		58		60	
R^2	0.52	0.62	0.16	0.30	0.07	0.20
ΔR^2		**0.11**		**0.14**		**0.13**

Stat.App. Table 3.. *N's* refer to number of teams across pooled seasons. Variables are standardized by season to confine model to estimates of the impact of predictors on variance in runs allowed within seasons as opposed to variance across them and to remove the effect of inter-season variability unrelated to the impact of the predictors on the outcome variable (Schell 1999, 2005). Beta *t*-statistics are in parentheses. Bolded predictors and ΔR^2s are significant at $p < 0.05$..

This could be a result of measurement-error disadvantages faced by DRA relative to DER and TZR. Like those systems, DRA estimates runs saved on the basis of the proportion of balls hit in play that are turned into outs by fielders at different positions. But unlike DER and TZR, DRA does not derive such information directly from Retrosheet reports. Instead, DRA imputes fielding opportunities to infielders based on league-wide base rates conditioned on various factors. For example, the number of fielding assists players at a position can be expected to make based on league averages are adjusted, first, for the proportion of team innings they played at that position; second, for their teams' rates of balls-hit-in play rate; and third, for the proportion of right- and left-handers on the their teams' pitching staff. The last of these factors is presumed to influence the resulting proportion of right- and left-handed hitters that batted against a team, and thus the proportion of balls hit in play to one or the other side of the infield. DRA likewise uses base rate estimates of the proportions of unassisted outs by first baseman that reflect popups and ground balls, respectively. While based on reasonable assumptions and validly converted into corresponding regression coefficients, this process necessarily multiplies the attenuating effects of measurement error at every step. There is no doubt measurement error associated with relying on the explicit reporting of the type and location of batted balls in Retrosheet event summaries, too, but it is likely to be much smaller than that associated with estimated fielding opportunities inferred from a series of base rates.

This is by no means to say that more could not be learned from wider use of DRA in empirical research. DRA is a serious and important effort to measure the contribution of fielding to runs avoided. The R^2s for models reported in this paper effectively rule out the possibility that the use of any alternative valid measure would modify the paper's central conclusions on the declining importance of fielding since the ascendency of today's unprecedentedly high strikeout and home-run rates. Nevertheless, it seems certain that were DRA runs-saved data made generally available for AL and NL players and teams, researchers could learn significantly more about the impact of fielding generally and about the quality of individual fielders in particular over the history of major league baseball. Such a development would indeed be a welcome one.

Supplemental Information References

Baseball Reference (undated). Position Player WAR Calculations and Details. Available at: https://www.baseball-reference.com/about/war_explained_position.shtml.

Heaphy LA (2003) *The Negro leagues, 1869-1960.* Jefferson, N.C.: McFarland & Co.

Humphreys MA (2011) *Wizardry : baseball's all-time greatest fielders revealed.* Oxford ; New York: Oxford University Press.

Schell MJ (1999) *Baseball's All-time Best Hitters: How Statistics Can Level the Playing Field.* Princeton University Press.

Schell MJ (2005) *Baseball's All-time Best Sluggers: Adjusted Batting Performance from Strikeouts to Home Runs.* Princeton University Press.