Fielding Before and After Baseball's Great Transformation

Supplemental Information

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 model 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 (R2_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 4 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 5 and 6 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 basemen 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 that 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 SI 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)	(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**

SI 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 offset run-environment scaling effects and 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 $p < 0.01$.

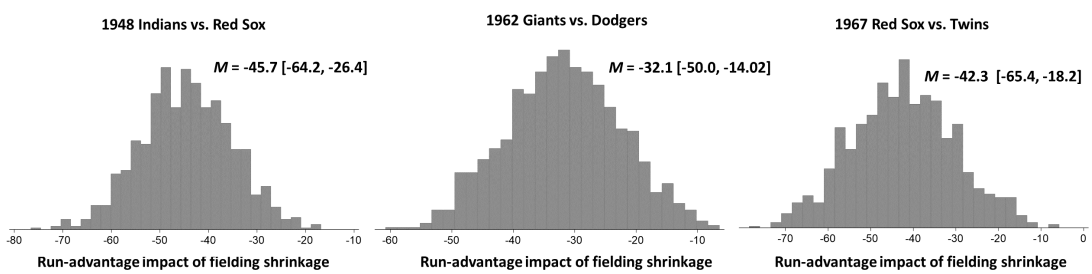
Whereas the single-season regression models use raw or untransformed values for the output and predictor variables, the multi-season regressions here use season-standardized values of the same. This transformation puts the predictors and explanatory variables for all seasons on a common scale, thereby 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

	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

SI 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 $p < 0.01$.

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



SI Fig. 1. "What if" Monte Carlo simulation outputs. Regression-model parameters (SI 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 typically fielded 10 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 60-80 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 SI 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**	**0.47**	**0.34**
	(6.90)	(5.17)	(3.23)	(2.47)	(5.40)	(4.05)
z_rfield		**-0.37**		**-0.39**		**-0.40**
		(-3.47)		(-3.35)		(-4.73)
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)
N	46		58		106	
R^2	0.52	0.62	0.16	0.30	0.22	0.36
ΔR^2		**0.11**		**0.14**		**0.14**

SI Table 3. *N's* refer to number of teams across pooled seasons. Variables are standardized by season to offset run-environment scaling effects and 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; 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. 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.