SOLVING DIPS

By Erik Allen, Arvin Hsu, Tom/Tangotiger, et al

The following is a compilation of various posts found at
http://www.baseballprimer.com/studies/archives/00000084.shtml

Interested readers are directed to go to that site for the full transcript of all posts. A 3 or 4 page summary of what appears below will be presented in a few days.

TABLE OF CONTENTS

PART 1 – Year-to-year correlations, and the meaning of “r”

Tom presented data on year-to-year correlations

Treating this as a first step, and realizing that we have parks, and switching teams to take into account, here are the year-to-year correlations of all 1687 pitchers with at least 500 PA in consecutive years (which itself may imply a selective sampling issue) from 1972-1992:

Event	r
K	0.78
BB	0.66
1B	0.47
HR	0.34
XBH	0.26
.....	
1bBIP	0.25
xbhBIP	0.21

Except for BIP, all are based on a per PA basis. XBH = (2b+3b)/PA.
1bBIP=1b/(PA-HR-BB-K), xbhBIP=(2b+3b)/(PA-HR-BB-K)

Later on we will find out the the year-to-year correlation (r) actually does not mean persistence of skill. We’ll get to that later.

Tom then presented similar data, but using a lower threshhold.

Lowering the bar to at least 250 PA in consecutive years, and you get the same order of results (all r are about .05 less than the above ones).

Event r
K 0.74
BB 0.61

1B 0.40
HR 0.30
XBH 0.22

1bBIP 0.18
2bBIP 0.17

Addressing the issue of DIPS, Tom said:

As for what it says of DIPS, there's no change. The year-to-year r is .20 as has been reported by many people many times with very different data sets. It's still our best guess that if a pitcher has a (1b+2b+3b)/BIP rate of .320 and the league is .300, then a pitcher's "true" talent, based on the BIP, is .305 (80% regression towards the mean, or 1-r). This applies to pitchers with 500 to 1200 PAs.

Addressing issues on GB and FB pitchers, Tom presented:

I broke up the pitchers with at least 250 PA in both years into "GB" and "FB" pitchers.

I ran the correlation only for the xbhBIP category. The FB pitcher's year-to-year r was .10, while it was .19 for the GB pitchers. Seems to me that park and OF fielders play a big part here.

Ok... I'll do the same for 1bBIP: .13 for GB pitchers, and .15 for FB pitchers. Again, makes sense.

Sorry, but I don't have the breakdown by FBhits, GBhits, though that would be very useful.

Erik Allen makes one of his first of many appearances by trying to get a meaning behind “r”.

While I agree with you that, in a strict sense, comparing correlation coefficients of two statistics from year-to-year is technically meaningless, I think in certain cases it can be useful. In McCracken's original DIPS work, he simply shows that there is simply MUCH less predictability in BABIP than in K/9, BB/9 etc. So much less, in fact, that sample size issues are probably not the sole cause. This discovery in itself is quite interesting, because it explains in some sense why it is difficult to predict pitcher ERA from one year to the next.

I think the larger problem is assigning a cause to this type of study. McCracken attributes the discrepancy in correlation coefficients to a COMPLETE lack of pitcher "control." Subsequent writing on this site and others has shown this to probably be false.

Summing up (wait, I actually had a point? :)) : I think that comparing correlation coefficients is a pretty rough test to use, and owing to sample size effects, and the old aphorism that "causation does not imply correlation," it is really difficult to show an effect exists and to attribute a reason to that effect.

Tom tries to reiterate what we are after

We are not trying to establish if a pitcher has a skill, even though I and others are saying that, when we look at the year-to-year correlation.

What we are really saying is "does this particular metric correlate well year-to-year.... and if it does NOT correlate well year-to-year, then we should not be using it as a basis to predict the next year's metric".

So, if we replace the "ability" talk with the "metric's persistence", I think we'd be more accurate.

So, regardless of the extent to which a pitcher has a skill at preventing hits on balls in park, we are saying that:

[Official quote]

the metric "hits per ball in park" has an r of about .20 among pitchers with 500 to 1200 PA, and therefore we need to regress that metric heavily (80% for the group, which may not necessarily apply to the individuals to the same figure), if you want to predict next year's metric.

Even having next year's metric still does not tell you about the pitcher's true underlying skill at preventing hits on balls in park. Just to the extent that we can measure this underlying skill, that's our best guess as to the expected outcome of that skill, with a [insert number] margin of error.

It may very well be that if we look at very specific breakdowns by zones, opponent, fielders, park, weather, etc, that we CAN ascertain what a pitcher's skill is at preventing hits on balls in play (see: PZR). It's just that, for the moment, the metric called "hits per ball in park" does not do a good enough job at establishing the pitcher's skill with "hits per ball in park". (This would be similar to ERA, earned runs per 9 innings, does not do a good enough job to establish a pitcher's skill at allowing "earned runs per 9 innings".)

[End Official Quote]

After a long discussion, and some test results produced by Erik Allen, Tom recaps as to what the year-to-year “r” really means

To recap, the year-to-year r is dependent on:
1 - how many pitchers in the sample
2 - how many PAs per pitcher in year 1
3 - how many PAs per pitcher in year 2
4 - how much spread in the true rates there are among pitchers (expressed probably as a standard deviation)
5 - possibly how close the true rate is to .5
6 - the true rate being the same in year 1 and year 2

Given all that, the biggest factor in the K "r" being the highest and the XBH "r" being the lowest may be entirely due to #4. That is, the "r" is not explaining #6 anywhere near as much as we think it is.

I think, maybe, that simply the tightness of the h-hr / BIP (over a career) is what is being explained and not the "persistence" of ability, based on the "r".

For those of us hoping that "r" was trying to find the signal, that's not what it's doing. The h-hr / BIP is too tight to find a signal.

So, we should use a heavily regressed h-hr / BIP, but not for the reason of "lack of control".

Bob Mong adds his view

However, Tango was right when he said that the numerator doesn't matter, only the denominator. The sample size matters, not the number of successes.

And furthermore, Tango was also right when he wrote that the closer you are to 0.5, the larger the standard deviation. That follows from the formula:

p, $p \times (p - 1)$
0.1, 0.09
0.2, 0.16
0.3, 0.21
0.4, 0.24
0.5, 0.25
0.6, 0.24
0.7, 0.21
0.8, 0.16
0.9, 0.09

As the probability gets further from 0.5, the standard deviation will become smaller, given identical sample-sizes. That is why the standard deviation of the out is smaller than the STD of 1B, and that is why the STD of XBH is smaller than 1B.

Erik adds to the summary on this subtopic

I think the basic lesson we can take from this discussion is: Year-to-year correlation coefficients depend on a lot of factors, including sample size, how often the event (e.g. hit) occurs, and the spread of talent. Therefore, a low correlation coefficient ON ITS OWN is not enough to say that a talent or persistence of ability does not exist. In fact, as the simple simulations I did above show, you can get a VERY low correlation coefficient even when a distinct talent is present.

PART 2 – Simulations to establish "true" rates that would produce "observed" rates

Tom makes an innocent suggestion, in which Erik will take and run with it far beyond what we were waiting for:

I would think that you create a model where you have known fixed talents, with a range equivalent to what you think MLB has (however you do that, but you can try different reasonable scenarios). And figure out the year-to-year "r" based on this model, and the number of BIP these pitchers have. That essentially gives you the "upper boundary" of r, which may be something like .2 or .25 for hits on BIP.

If in actual life, the MLB r is .18, well then, that's pretty strong evidence of perisistence, right?

Erik posts the BABIP (batting average on balls in park)

#BIP BABIP
100-199 0.291
200-299 0.284
300-399 0.284
400-499 0.282
500-599 0.282
600-699 0.278
700-799 0.275
800-899 0.272
900-999 0.268

And Erik continues his presentation

First a little background: As tango points out in his previous posts (56,57), in the PA apperance range of 200-800 BIP, the BABIP rate is very tight (0.275-0.284). I am therefore making the approximation that the pitchers that appear in this sample have a distribution of talent at preventing hits on balls in play. I am furthermore assuming that this distribution of talent is normally distributed (i.e. bell curve shaped) about 0.281.

Here are the statistics I get for all pitcher seasons between 200 BIP and 799 BIP (based on data provided by tango):
# of seasons: 4389
average BABIP: 0.281
Standard Deviation of BABIP: 0.027

The standard deviation is a measure of the spread of the data. Basically, we can say that 67% of all seasons should be between +/- 1SD of average (0.254 - 0.308) and 95% will be within 2SD. Nothing exciting here, this has all been done before.

The question that we cannot answer from the basic analysis above is: what is the standard deviation of "true" talent? For example, do all pitchers simply have the same talent level of 0.281 BABIP? Or, is there some spread to pitcher talent? What is the magnitude of this spread?

To answer these questions, we have to account for the number of trials (i.e. the number of balls in play). So, I have broken down the pitching seasons by balls in play into groups of 100. Listed below are the number of seasons in each group, and the standard deviation of the group:

#BIP	#seasons	STDEV
200-299	1446	0.032
300-399	812	0.0268
400-499	592	0.0245
500-599	507	0.0221
600-699	579	0.0210
700-799	454	0.0204

From the above table, you can see a clearly decreasing trend in the standard deviation of BABIP as you increase the number of BIP. And, intuitively, we can agree with this idea. After all, in a small number of trials, any number of fluky things can happen, including having a 0.400 BABIP or a 0.150 BABIP. As you increase the number of chances, the likelihood of a really fluky season decreases.

We now have standard deviations of the OBSERVED data broken down by number of balls in play. However, we also know that this observed standard deviation is not equal to the standard deviation of the TRUE talent level. For example, if all pitchers have the same inherent skill level (BABIP=0.281) the stdev of the true distribution is 0. The observed stdev will be something greater than zero.

To figure out what the true standard deviation is that matches the data, we can run a simulation. In this simulation, I set the true standard deviation of the group of pitchers, and measure the output observed standard deviation. Then, I tinker with the set value of the true standard deviation until the output standard deviation I obtain is equal to the observed standard deviation of the group. I can do this for the various numbers of balls in play.

This is already getting long, and I am probably rambling incoherently, so let me simply get to the data. The table below lists the number of balls in play range, and the TRUE standard deviation that would lead to the OBSERVED standard deviation given in tango's data.

#BIP	TRUE STDEV
200-299	0.014
300-399	0.014
400-499	0.012
500-599	0.012
600-699	0.012
700-799	0.012

I was ecstatic, to say the least. What I see above is a remarkably consistent picture of pitcher ability. It seems that, as a rough estimate, we can say that pitcher abilities are normally distributed about BABIP 0.281 with a standard deviation of 0.012 or so.

Roughly 2/3 of all pitchers should have a TRUE BABIP rate of 0.269-0.293. Roughly 95% of pitchers should have a true BABIP rate of 0.257-0.305. If this stands up, it is useful because it means that when a pitcher has a season with a 0.250BABIP, we can hypothetically give an estimate of his TRUE BABIP rate.

There are a ton of holes that can be poked in this, being as rough a calculation as it is, and I welcome any and all comments.

Tom chimes in and mentions we should consider 2 important variables

2 things I forgot about: park and fielding.

Erik reaffirms what we have to think about

Tango, your comments in 61 make me realize that I was getting ahead of myself. All the simulation above tells us is that we can match the observed experimental distribution if pitcher BABIP rates are normally distributed with a stdev of 0.012. We have not yet made any claims as to why this distribution of true BABIP rates exists...is it due to the pitcher, the defense, or the park? This is obviously a key question in predicting abilities going forward.

Tom offers ideas for the 2 variables

To simulate park, that's easy enough. Just go to the above link. We see that the stdev for park is .0085. Since they play half their games at home, the "seasonal" park adjustment would be .004.

We definitely have to simulate fielding, but the question is "how"? If I look at team-level UZR, on a year-by-year basis (n=120 over 4 years), the stdev is about .0100 (but you need to regress somewhat). If I take it on a multi-year basis (1999-2002, n=30), the stdev is .0070. Since teams do turnover, I think the answer lies somewhere in-between, I'd guess. So, I'd make that .008. (I'd guess that if you even just used ZR, or any other measure, you'll get similar results.)

If you were to run your simulation where you set the standard deviation of the park to .004 and the fielders to .008, we can figure out what's left over for the pitchers.

Now, you can try running your sim so that fielding is set to .006 or .010 or anything (reasonable) you want really. So, you can say that "if fielding stdev is .006, pitching stdev is .007... if fielding stdev is .008, pitching stdev is .005", or something alone those lines.

This is really exciting! We can finally come up with the proper "split" between fielding, pitching, and park.

Erik posts more results

Before getting into more complicated simulations, I thought it would be appropriate to first look at some "extreme" cases.

The first question one might ask is: Is the pitcher ENTIRELY responsible? The answer is almost certainly no, but it might be instructive to see what kind of results you would expect if such was the case.

What I did in this set of simulations then, was to randomly assign 10,000 pitchers a BABIP skill level (for example, 0.281, or 0.250, etc.). These skill levels are normally distributed about 0.281 with a standard deviation of 0.012 (as found previously to fit the data). We assume that this BABIP level is ALWAYS their true level. Then, I simulate 2 separate seasons, and record and OBSERVED BABIP level for each pitcher each season (this would correspond their true major league performance). I then measure the correlation coefficient for the year-over-year data, and compare it to the correlation coefficient tango found in his study (0.15, see post 6).

Okay, wordy, I know, so let's get to the data: In the data file tango sent me, there were 4389 pitchers that had between 200 and 800 BIP in a given season. The average number of BIP was around 430. Therefore, I let each pitcher have 430 BIP in each season.

#BIP; 430 for both seasons, for all pitchers. r = 0.24

As we would expect, the correlation coefficient is too large. One modification we could make to change the outcome slightly, would be to assign different pitchers different numbers of plate apperances, to more closely reflect reality. When I do this (e-mail me if you want more methodology), I get r = 0.21. Still too large.

The "Well, duh!" conclusion, is that pitchers BABIP talent does not lie solely with the pitcher.

A second extreme case would be to ask if the data can be explained solely on the basis of park factors. That is, the pitching has no influence, the defense has no influence, only the effect of the park determines the BABIP rate.

Tango has a list of BABIP park factors at his website (see homepage link). You divide these factors by 2 to get a team's park effect over the course of a season. The standard deviation of this distribution is 0.004.

In 2002, the average team had around 4550 BIP over the course of a season. Using the same methodology as above, I assign each team a BABIP level based on a normal distribution with standard deviation 0.004.

The correlation coefficient for this case is r = 0.25.
The year-to-year correlation coefficient on a team level is more like r = 0.6 So, clearly, the park is not the only factor either.

Erik runs his first sim with a random park for a random pitcher

I ran the simulations that tango suggested. That is, I introduced a random, normally distributed defensive factor for each player. Tango sets the standard deviation of the defensive contribution at 0.008. However, since I didn't know the exact basis for this number, I ran the simulation under 2 assumptions:

Case 1:
Assumptions:
1. 0.008 is the _observed_ stdev of defensive talent, AFTER ACCOUNTING FOR PARK EFFECTS.
2. The talent of a defense is independent of the park they play in (i.e. the park effect and the defensive ability of the team are independent variables).

In Case 1, we need to determine the true standard deviation of defensive ability, since the observed standard deviation is larger than the true standard deviation. To do so, I ran my simulation at different levels of true standard deviation, and measured the output stdev (each team was given 4550 BIP). I get a true standard deviation of 0.0045.

After doing this, I can compute the true standard deviation of pitcher ability. I use the same tactic as in previous posts, changing the true stdev to match the observed stdev for different levels of BIP. The stdev of pitchers in Case 1 is 0.010. Table 1 presents some data I get from such an analysis:

#BIP	Simulation_stdev	Real_life_stdev
250	0.0308	0.0321
350	0.0266	0.0269
450	0.0241	0.0245
550	0.0225	0.0221
650	0.0211	0.0210
750	0.0202	0.0204

As you can see, the simulations stdev matches the "real life" stdev for every case expect for the pitchers in the 250 BIP range. This is a problem I have been having fairly consistently. I think it could be explained by a number of factors:
1. 100 BIP is too wide a range to use at such a low number of BIP
2. The range of talent for pitchers in this group is larger

Anyway, for case 1, we see the pitching/defense/park breakdown is 0.010/0.0045/0.004

P.S. I also calculated a correlation coefficient as described previously (pitchers assigned 200-800 BIPs according to the major league distribution). I get r = 0.20. Still too high (as expected)

Case 2:
Assumptions
1. 0.008 is the true distribution of defensive talent
2. Defensive talent is independen of park

The only difference from case 1 is that I now use 0.008 for the stdev of defensive talent. Using the same procedures as above, I get a pitcher stdev of 0.007. See table for simulation details:

#BIP	Simulation_stdev	Real_life_stdev
250	0.0306	0.0321

```
350 0.0266 0.0269
450 0.0240 0.0245
550 0.0222 0.0221
650 0.0209 0.0210
750 0.0199 0.0204
```

Case 2, the pitching/fielding/park breakdown is 0.007/0.008/0.004

P.S. I also calculated a correlation coefficient as described previously (pitchers assigned 200-800 BIPs according to the major league distribution). I get r = 0.19. Still too high (as expected)

PART 3 – Doing statistically what was done as a simulation

Arvin foreshadows an important event

I need to think about this. Chris, this may be where you had been planning to go. I figure we should be able to calculate, rather than simulate what our sample variance _should_ be. This would be an "exact" formula.

Arvin adds more statistical insight

What does this mean for us? It means as N increases, the sample std-dev of the sample will decrease. This explains, perfectly, Erik's findings in post #58:
BIP #seasons std(p-hat)
200-299 1446 0.032
300-399 812 0.0268
400-499 592 0.0245
500-599 507 0.0221

to take two numbers:
n=250, std(p-hat)= .032
n=500, std(p-hat) = std(p-hat(n=250))/sqrt(2) = .032/1.41 = .0226
You're std(p-hat) for n=550 is .0221!!!

Tom starts publishing some fielding standard deviations

Both: .009
IF: .013
OF: .013

rf: .020
2b: .020
ss: .021
lf: .022
cf: .026

3b: .031
1b: .032

(Doing a weighted average of the above, and we get a value of .024. I think for ease, we should consider the standard deviation on a per-position basis to be the same and equal to .024. Erik, it's your time, so do whatever you figure you can handle.)

These standard deviations are all observed and need to be sim-ed or calculated to determine the "true rates".

Arvin finally provides the equations to match the simulations

As for formulas... I think the numbers you've been simulating, Erik, can be approximated by assuming that the variance's add.
IOW,
k ~ Bin(n,p)
p ~ Norm(0,sigma^2)
Var(k) = n*p*(1-p)
Var(p-hat) = p*(1-p)/n <---- this is what we expect the Binomial to contribute to our data.
Observed Variance of Data = Var(p) + Var(p-hat)
= sigma^2 + p*(1-p)/n

So... Using your data: sqrt(p*(1-p)/n+.012^2)
200-299 1446 0.032: .0309
300-399 812 0.0268: .0269
400-499 592 0.0245: .0244
500-599 507 0.0221 .0226
600-699 579 0.0210 .0213
700-799 454 0.0204 .0203

And he makes perhaps the most important declaration to date

Also, if you have multiple sources of variance, they will also add similarly:

Total True Variance = True Defense Variance+TrueParkVariance+TruePitcherVariance

.012^2 ~= .0075^2 + .008^2 + .004^2 = .0117^2

I'm not entirely comfortable with the simplification that the Norm Variance and the Bin Variance will linearly add, but it appears to fit the data well.

Using your data: sqrt(.0075^2+.008^2+.004^2) = .012

Tom tries to apply this new equation

Arvin's theorem is intriguing. For example, I mentioned that the observed stdev for IF and OF was .013, and for the team it's .009 (according to my post 111).

Let's see what happens with this new equation, and realizing that half the BIP are IF and half are OF (let's say).

Observed team ^ 2 = [(.013/2) ^ 2] + [(.013/2)^2] = .009 ^ 2

Wow!

How about if we use the .024 for each of the 7 positions? Following the same process, and we get: .009!

Holy moley!

Now, if you want to really impress me, tell me how to get from the observed stdev to the true stdev. That is, how much do I regress towards the mean, given the sample size? Do I make it k/sqrt(n)? How do I know what to set k to?

And Arvin responds

Observed Variance = Binomial Variance + True Variance

So...

(Obs. Std)^2 = p*(1-p)/n + (true Std)^2

And this is what the chart that I posted after that showed for Erik's numbers. Oh, p is the avg. rate for the binomial: eg. .281, but it would be different if the team fielded at, say, .300, or whatever...
And n is obviously the number of BIP.

Tom applies the equation

So, we have

.0090^2 = .28*.72/4500 + true^2

that makes the true std dev at the team level as: .006

So, that's the fielding.

.012 ^2 = .006^2 + .004^2 + pitching^2

pitching = .010

So, are we saying that each pitcher has a .010 stdev, each team of fielders is .006?

Tom concludes

So, what we are saying is that we have a 10/6/4 split between pitching/fielding/park, in that order. Luck plays a part, and that is dependent on the sample size. When n=1, it's almost all luck. When n=1 million, luck is not involved.

So, over 700 BIP, where we observed a .020, we have the following:

observed ^ 2 = .010 ^ 2 + .006 ^ 2 + .004 ^2 + luck ^2 = .020 ^ 2
solving for luck = .016

So, can we say that when a starter has 700 BIP, the influence on those BIP as a group can be broken down by:
luck : 44%
pitch: 28%
field: 17%
park : 11%

I have to admit that I've recently said, though I don't remember where, that I thought the split would be 40/30/20/10 with the order being luck,fielding,pitching,park.

What we are saying here is that pitching and not fielding is the larger determinant between the two. And perhaps before I read about DIPS I might have had the correct order.

I think it's still important that yes we need to separate the components (HR.BB,K) from the BIP, as Voros does. But, the conclusions drawn from that does not stand based on the reasoning.

I think our best conclusions would be the follows:
1 - pitching has more impact on BIP than does fielding
2 - luck has more impact than anything, over 700 BIP
3 - BABIP is not a good enough measure for the pitcher's skill

What would be interesting is that if MGL or Tippett or someone with pbp data gets around to implementing the PZR blueprint I published (the flip side to UZR), that we'll get closure on this subject. That is, we should be able to get the standard deviations on the pitcher's side that will support the data we are inferring here.

So, before we trample in any direction, it may be worthwhile to keep the case open, pending final data. After all, we may have made a serious miscalc somewhere.

I'm not really sure of the impact. It's still a blur
to me as to what use to make of it.

What we are saying is that there are 2 components for a pitcher: his non-fielding dependent skill (HR.BB.SO) and his fielding-dependent skill.

We know very well how to estimate the former, and not very well the latter. Since the BABIP figure is not reliable for an individual pitcher, it's more accurate to use say 50% lg, 40% team, 10% pitcher to estimate his expected BABIP. But, that estimate will come with a very wide margin for error.

The conclusion stands that you need to separate things, and you can't rely on a pitcher's past BABIP to predict the future (much like you wouldn't use his ERA). Still outstanding is WHAT to use for BABIP. I'll contend that PZR would be that measure. But, that has yet to be implemented by anyone.

Arvin concludes as well

Or... you could just say:
Binomial distribution, n=700, p = .281
std: sqrt(p*(1-p)/n) = .0169

And Tom affirms

Actually, the observed should have been
600-699 579 0.0210
700-799 454 0.0204

So, at 700 BIP, I should have used .0207. Reworking, and we get a nearly perfect match.

Matt Goff confirms a previous statement by Arvin

> I think it comes down to why we're adding variances:
> In a normal distribution, you add when you add two random variables.
> eg. X ~ Norm(3,.05^2), Y ~ Norm(4,.03^2)
> Z = X+Y
> Z ~ (7,.05^2+.03^2)
>
> Although I'm not yet convinced it works this way for a normal p affecting a binomial, the sims seem to bear it out. Why? Well it makes sense. On some level, you're adding two r.v.'s, the binomial
> and the normal.

It seems to me that your concern about the addition of variances may be alleviated by recalling that a binomial is approximately normal for large n (thanks to the good old Central Limit Theorem). I can't remember the exact rule of thumb, but it seems like n*p>25 or something like that. In any case, if I understand what I have been reading, the n should be quite adequate for the normal approximation to hold.

Tom adds one last reminder

I just want to interject something to keep in mind. Remember, our equation is

trueDER ^ 2 = truePitch^2 + trueField^2 + truePark^2

Erik has provided trueDER from his sim, and Arvin has confirmed it with his "observed" equation, and that is .012. I've provided the truePark figure as .004. Dropping all the decimals, and our equation becomes

128 = truePitch^2 + trueField^2

Based on UZR, which I'll have to go over because I'm not sure I'm using the right numerator (Levitt's numbers might include HR), the observed single position UZR is around .025 and the observed team fielding UZR is around .010. So, our true UZR will be somewhere between .005 and .008, probably.

We're not even sure that UZR is the best thing to use, but it is the best thing available at the moment. (You could even use ZR, and I'm pretty sure you'll get a single position observed stdev of .025 for your regular players. This is easy to eyeball since the range of players is mostly around +/-.05 outs/BIP, so that would be 2 standard deviations.)

Anyway, so we've got something like
128 = truePitching^2 + [5 to 8]^2

So, when fielding = 8, pitching = 8.
When fielding = 5, pitching = 10

etc, etc.

So, depending on how the fielding measure is determined and manipulated, a small change there will have a huge impact in the relative value between fielding and pitching.