Journal of Quantitative Analysis in Sports

Volume 7, *Issue* 3 2011 *Article* 7

You're Hurting My Game: Lineup Protection and Injuries in Major League Baseball

David C. Phillips, *Georgetown University*

Recommended Citation:
Phillips, David C. (2011) "You're Hurting My Game: Lineup Protection and Injuries in Major League Baseball," *Journal of Quantitative Analysis in Sports*: Vol. 7: Iss. 3, Article 7.
Available at: http://www.bepress.com/jqas/vol7/iss3/7
DOI: 10.2202/1559-0410.1296

You're Hurting My Game: Lineup Protection and Injuries in Major League Baseball

David C. Phillips

Abstract

Does lineup protection exist? Observers of baseball frequently invoke the idea of lineup protection in which a batter will walk less and hit for more power if he is followed in the batting order by a high-quality hitter. Previous attempts to measure protection in Major League Baseball have sometimes found evidence of lower walk rates but never an impact on power hitting. I argue that these efforts fail to uncover such evidence because lineups are selectively chosen by managers, introducing endogeneity bias into ordinary linear regression and batting split comparisons. To remedy this problem, I use injuries to batters' "protectors" as a natural experiment that quasi-randomly changes the level of protection received by batters. Using this approach, I find evidence that protected batters hit for more power, hitting 9.7 percent more extra-base hits if the protector's OPS is 100 points higher. These effects are strongest among 3rd hitters, who hit 26 percent more extra-base hits under the same scenario. Supporting the previous literature, I find that batters walk more, especially intentionally, when left unprotected.

KEYWORDS: lineup protection, natural experiment, optimal lineup

Author Notes: I am deeply indebted to the sabermetric community that generates such extensive, open-access data. In particular, this study would not be possible without the contribution of Josh Hermsmeyer in creating a publicly available injury database. Also, this study relies heavily on Retrosheet play-by-play data; many thanks to everyone who has contributed to Retrosheet. I have also benefitted from the comments of Guy Molyneux, Phil Birnbaum, and an anonymous reviewer. Of course, all errors and omissions remain my responsibility.

[Manager Jim Riggleman] flip-flopped Ryan Zimmerman and Adam Dunn, moving Dunn up to third and hitting Zimmerman in cleanup. 'I wanted Zimmerman hitting behind Dunn today and Willingham hitting behind Zimmerman,' Riggleman said. 'Try to protect them the best I can. We're struggling a little bit to score runs.' [1]

1 Introduction

Few ideas influence the choice and discussion of batting order in Major League Baseball more than the notion of 'protection.' According to this theory, if the current batter is followed by a poor hitter in the batting order, then the pitcher will often 'nibble around the edges,' i.e. throw pitches further from the center of the strike zone. This results in pitches that are more difficult to hit sharply but also more likely to be called balls. Ultimately, this theory predicts that an unprotected batter will be walked more frequently but will not hit the ball with as much power: fewer doubles, triples, and homeruns. For the purposes of this paper, I will frequently refer to this strong form of the protection theory as just 'protection.' A weaker but similar theory holds only that batters will be walked more, often intentionally in very particular situations, if left unprotected. I will refer to this as 'weak protection.'

The above quote provides an example of two important facts. First, baseball players, managers, and commentators almost universally believe in protection and frequently invoke it, as Washington Nationals manager Jim Riggleman does in the quote above. Second, managers make decisions regarding batting order based on both performance and protection. Because Dunn had started slowly in April and early May of the 2010 season, Riggleman occasionally moved number 4 hitter Adam Dunn out of a 'protector' role behind one of the best (if not the best) Nationals hitters, Ryan Zimmerman, and into a place of protection in front of Zimmerman.

Despite being widely believed by practitioners of baseball, the protection theory has not been supported by statistical studies to date. In fact, an iconoclastic disbelief in protection has become common among those who use statistics to analyze baseball. This literature can be traced back to a study by Bill James (1985) of whether the batting performance of Atlanta Brave Dale Murphy stagnated when his protection, Bob Horner, was injured during the 1976-1984 seasons.[2] He found no evidence of such an effect. The focus on

[1] Duffey (2010)

[2] The literature generally credits this study to James, though James (1985) actually credits the original analysis to Jim Baker. I will follow the literature in referring to this as

only one player, of course, is open to criticisms of external validity: maybe Dale Murphy is an unusual player who can rise above the need for protection. Additionally, comparisons of batting splits from a small number of plate appearances can be subject to very large uncertainty, something James later called into question (James (2005)). Finally, even the conventional standard errors used to measure uncertainty in such studies may be too small, because they require the assumption that batting outcomes of two plate appearances are independent. This is unlikely to be true due to ballpark factors, opponents, weather, and potential 'streakiness' of hitters.[3] If any of these factors are excluded from the model in question, the standard errors will be too small.

To address some of these issues, subsequent studies have expanded the scope of players and/or seasons used to test the protection theory, generally confirming James' result of non-existent or miniscule effects of protection. Because they use more data and are more comprehensive in player coverage, these studies make great strides toward better generalizability of the results and more precise measurement. However, these studies differ from James' original study in a subtle but important way: while James' study focused on a case where changes in protection were caused by injuries, these studies examine changes in protection caused by a myriad of factors. Unfortunately for the statistician, managers frequently change the batting order in response to a player's past or expected performance, as demonstrated in the quote above. As a result, protection is endogenous because there is likely to be some third variable that affects the protection received by the batter while also having a direct impact on performance. In fact, there are many such variables, and some previous studies control for a great many, including player quality and the state of the game. However, it is rare that we will ever be able to observe all of these factors. For example, if a manager sees that a star player has been performing very poorly, he will likely move the batter down in the lineup to a position of less protection. Given the large amount of randomness in baseball, that hitter's performance is likely to improve as it regresses toward his actual skill level. As a result, a player may actually be observed to perform better when unprotected, even if the causal impact of protection is positive.

To avoid this selection bias one would prefer to analyze protection in an environment where lineups are not selectively chosen by managers. This highlights the power of Bill James' original idea of looking at changes in protection primarily brought about by injuries. Holding age and a few other observable

James' study of protection.

[3]The idea that hitters can be 'hot' or cold, i.e. batting outcomes are positively serially correlated, is another idea widely believed in baseball but debated by those that study the statistics.

DOI: 10.2202/1559-0410.1296

factors constant, the timing of an injury is largely unchosen and random. As a result, injuries to a player that serves as a 'protector' provide a random, exogenous change in the quality of protection. They provide a natural experiment in which changes in batting outcomes can be plausibly attributed to the causal impact of protection, rather than the myriad of other factors influencing the choice of batting order.

In this paper, I study batting outcomes of hitters on all plays in Major League Baseball (MLB) from 2002-2009. By combining Retrosheet data on batting outcomes with a recently compiled database of disabled list trips for all MLB players, I am able to extend Bill James' original idea and study how batting outcomes respond to injuries of a batter's protection for all players from 2002-2009. In this data, I find support for the strong version of the protection hypothesis. Injuries to a batter's protection cause a measure of quality of the on-deck hitter, onbase plus slugging (OPS), to fall by 28 points. Over the whole sample, this drop in protection leads to batters receiving 26 percent more intentional walks and hitting 3 percent fewer extra-base hits. The magnitude of these effects indicates they are of noticeable importance to batting outcomes, even for a modest change in protection. Also, given that there is no change found in the total number of hits, this indicates that batters are indeed hitting with less power, not simply putting the ball into play fewer times.

If the protection hypothesis is true, we would expect these effects to be strengthened in the classic case of protection: elite hitters batting in front of other elite hitters. To check this, I also analyze the results among hitters who usually bat third in their lineups. Among these hitters, the results are even stronger, with 3rd batters seeing statistically significant differences in all relevant outcomes in the direction predicted by protection. Third batters walk more (both intentionally and in general) and hit fewer doubles and homeruns when their protection is injured. Once again, hits in general show no difference. The effects are much stronger than for the whole sample with point estimates generally two to three times larger than for the whole sample. These results indicate that injured protection shifts production from power hitting into walks and singles, especially for the best batters.

To help interpret the results, I use instrumental variables analysis to translate the impact of injuries on outcomes into the impact of on-deck hitter quality (measure by OPS) on batting outcomes. Given roughly linear effects, a 100 point (one standard deviation) increase in on-deck OPS decreases the intentional walk rate by around 90 percent (from the mean) and increases the extra-base hit rate by almost 10 percent. The effects are even larger for third hitters who would see their extra-base hit rate increase by about 26 percent.

I also test the impact of protection on a measure of overall production but find no statistical evidence regarding this overall effect. The main two impacts of protection, fewer walks and more extra-base hits, affect overall production in opposing manners, and I do not have enough statistical power to measure which effect dominates in terms of overall production. Thus, I find evidence of protection affecting the distribution of batting outcomes, but I cannot make statistical statements either way on the overall impact of protection.

Finally, in all of the analysis I make use of cluster-robust standard errors, which correct standard errors for violations of the assumption that batting outcomes across different plate appearances are independent. These standard error estimates have become common in other fields but have not seen wide use in the statistical analysis of baseball. Given the low likelihood that at bats are actually independent events, statistical analysis of baseball could benefit from this tool.

The remainder of the paper is as follows: the second section explains the methodology of the paper and discusses the literature; sections three and four describe the data and the results, respectively; and section five concludes with a summary of the results and a discussion of the implications of this paper for further research.

2 Methodology and Literature Review

2.1 Endogenous Protection

To couch the protection hypothesis in a regression framework, I focus on an equation such as equation (1):

$$Y_{ibt} = \alpha + \beta P_{ibt} + \epsilon_{ibt} \quad (1)$$

where Y_{ibt} is a dummy variable indicating a particular batting outcome for plate appearance i of batter b during season t; P_{ibt} is a measure of quality of the on-deck hitter (protection); and ϵ_{ibt} is a random error term. As a linear regression on a dummy dependent variable, this can be interpreted as a linear probability model.[4] The protection hypothesis supposes that the coefficient β is positive if the outcome of interest is extra base hits and negative if the outcome of interest is walks and perhaps singles. The researcher approaches

[4]Linear probability models give very similar results to probit, logit, etc. but make instrumental variables and clustered standard errors much more straightforward. As such, I will use linear probability models at all times in this paper.

DOI: 10.2202/1559-0410.1296

this equation with a causal interpretation, that the sign of these coefficients reflects the causal impact of protection on a batter's outcomes. The problem, of course, is that the raw correlation between batting outcomes and the quality of protection is likely driven by other factors. Even absent protection we would expect a positive correlation due to the fact that managers selectively place their best hitters together near the top of the batting order. To control for hitter's ability as well as the way that game situation, opponents, etc. affect the batting outcome, we improve on equation (1) to consider equation (2).

$$Y_{ibt} = \alpha + \beta P_{ibt} + \delta X_{ibt} + u_{bt} + \epsilon_{ibt} \quad (2)$$

where X_{ibt} is a set of control variables and u_{bt} is a batter-year fixed effect (dummy variable). Importantly, the batter-year fixed effect controls for any variable that stays constant for a given batter across a season, including overall player ability, approximate age, etc. The impact of including the fixed effects becomes apparent if we take the average of equation (2) over all plate appearances in a batter-year and subtract it from (2). This gives an equivalent formulation:

$$Y_{ibt} - \bar{Y}_{bt} = \beta \left(P_{ibt} - \bar{P}_{bt} \right) + \delta \left(X_{ibt} - \bar{X}_{bt} \right) + \epsilon_{ibt} - \bar{\epsilon}_{bt} \quad (3)$$

where $\bar{P}_{bt}$ is the average protection across plate appearances in a batter-year; other variables with 'bars' on top are likewise defined; and the constant and fixed effects have disappeared because they are constant across plate appearances within a batter-year. First, this makes it clear that, despite some minor differences due to linear functional form and slightly different controls, the fixed-effects specification is very similar to that used in previous work that controls for averages of the dependent variable (e.g. Bradbury and Drinen (2008)). Second, it clarifies what variation in protection is being used to identify the impact of protection. In this specification, the variation that matters is the difference from a batter's mean level of protection in that season, i.e. how a batter's protection differs from the protection he typically has.[5] Given that for each on-deck hitter on-deck OPS is measured at the season level, these changes only occur when the lineup changes. Some changes in lineup, like injuries, are likely to be random and thus good experiments to tell us if protection impacts performance. However, most deviations of batting order from a team's typical

[5]Note that if we remove the control variables and change the protection variable from a continuous variable to a dummy variable that indicates protection if on-deck batter quality is above some threshold, then this formulation is also equivalent to comparing average batting lines when protected and unprotected, as many previous studies have done including James (1985), Grabiner (1991) and Tango, et al (2007)

lineup are the result of tinkering by the manager based on past or expected performance of the players. It is likely that many variables, unobservable to the researcher, affect this protection decision while also having a direct impact on batting outcomes. This violates the important regression assumption of exogeneity and induces correlation between protection and performance that cannot be interpreted as the causal impact of protection.

Consider an example: early in the 2010 season, the Chicago Cubs number four hitter, Aramis Ramirez, was hitting very poorly. In response, the Cubs moved him down to the fifth spot, and then about one month into the season a media report noted, '[Cubs manager Lou] Piniella tried to get Aramis Ramirez, now batting .154, back in the swing by switching him in the order with Alfonso Soriano, from fifth to sixth.'[6] In other words, due to his poor performance, Ramirez was moved to a position of even less protection, further down the order. Now, suppose that Ramirez is not 'cold' but simply suffering from some combination of ageing and poor luck. In a small sample, the luck factor is likely large (Tango, et. al. (2007)), and Ramirez's performance will almost certainly regress back toward his solid career numbers, at least partially.[7] As a result, he will probably perform better batting fifth and sixth than he did in the fourth spot. If this happens and the effect outweighs the impact of protection, we again would observe a negative correlation between the quality of his protection and his overall batting outcomes, even if protection exists.

Of course this is just an anecdotal example, and many more stories could be conjured. This is exactly the point. Managers choose their lineups based on a very large number of considerations, whether real or imagined. Given that these considerations are based on past performance, protection quality and performance numbers will be connected in many complicated and unobservable ways. In technical terms, protection is endogenous in the regression. Equivalently, there is always omitted variables bias. It is virtually impossible to include every relevant variable in the regression. Even if we could identify all of the important factors, many of them are simply unobservable.

2.2 The Experimental Solution

In the ideal situation, a researcher would overcome these issues by randomly assigning quality of protection. If one could randomly select a portion of a batter's plate appearances to receive protection from a star teammate and se-

[6]Muskat (2010).

[7]In fact, this has been realized since the original version of this article was submitted. Ramirez posted a .648 OPS during the first half of 2009 but an .847 OPS in the second half, right in line with his career OPS of .839.

DOI: 10.2202/1559-0410.1296

lect others to receive less protection from a weaker hitter, then any correlation between protection and batting outcomes could be attributed to the causal impact of protection. All of the previously discussed problems caused by selective choice of batting orders would be removed. This type of experimental impact evaluation has been common in medical and agricultural trials for many years and is increasingly used in the evaluation of economic and social programs. However, this method is not feasible to study lineup protection without the ability to intervene in lineup selection. Though seemingly feasible, I am not aware of such experimentation in baseball.

If a randomized trial is the ideal, then we would like to exploit a situation that creates variation in lineup protection in a manner that, if not completely random, is as good as random with respect to batting outcomes. Basically, we would like a variable with two qualities. First, it must have an impact on the level of protection received by batters. Second, it must not, after controlling for some observable variables, affect batting outcomes through any channel except protection. These two requirements can be referred to as 'relevance' and 'validity' while the variable induced by the natural experiment is an 'instrumental variable.'

In our context, injuries to a batter's protector provide exactly this sort of natural experiment. Using available data, I determine protection relationships as described in the data section. Then, the relevance of this natural experiment can easily be checked by regressing the actual on-deck hitter's OPS on injuries to the batter's protector(s).

$$P_{ibt} = \alpha + \beta I_{ibt} + \delta X_{ibt} + u_{bt} + \eta_{ibt} \quad (4)$$

where protection is being measured by the OPS of the hitter who actually follows batter b for plate appearance i and I_{ibt} is an indicator of whether any of the players designated as a protector of batter b during season t are injured during plate appearance i.

While relevance can be directly tested, validity of the instrument is difficult to test[8] and must be justified a priori. In our case, conditional on the age of the protector and the month of the season,[9] the timing of an injury

[8]Tests do exist, such as Hansen's J test, but it has two major problems. First, you must have more instruments than endogenous variables, which is often difficult to meet in practice. More importantly, the null hypothesis in these tests is that the instrument in exogenous, i.e. valid. As such, the tests can reject an instrument as poor but cannot confirm a good instrument.

[9]The approximate age of the protector, as well as any of his other characteristics that remain constant over a season, are absorbed by the batter-year fixed effect. This is because protectors, as defined here, do not vary during a season. Regarding controlling for months,

to a batter's protector is basically random and unrelated to batting outcomes except through the lower quality protection it causes. In other words, nothing about the protectee and his environment changes when the protector gets hurt except the level of protection he receives. This assumption provides the fundamental foundation of the analysis that follows. All regression analyses require an assumption similar to this one. In regression (3), the equivalent assumption is that all changes in lineup protection are unrelated to changes in batting outcomes, except through the causal channel of protection. As discussed above, this assumption seems implausible. Making a similar assumption about the relationship between protector injuries and batting outcomes seems much more plausible, since injuries result from unchosen and seemingly random events.

2.3 Reduced Form and Instrumental Variables

I will use two slightly different methods to analyze the data: reduced form and instrumental variables. Reduced form examines the direct impact of protector injuries on batting outcomes. The following equation, estimated by OLS, is the reduced form:

$$Y_{ibt} = \alpha + \gamma I_{ibt} + \delta X_{ibt} + u_{bt} + \epsilon_{ibt} \quad (5)$$

where variable definitions are the same as above. For interpretation, the parameter of interest is γ, which measures the change in batting outcome Y_{ibt} caused by an injury to a batter's protector. In this equation, I will include a full set of controls including a national league dummy, the pitcher's ERA, a platoon advantage dummy, out and inning dummies, dummies for each of the 8 possible combinations of runners, and ballpark dummies. Importantly, batter-year fixed effects (dummy variables) and month dummies will also be included. Month dummies are important to include because early in the season, there are few injuries and lower performance. As described above, batter-year fixed affects control for a batter's characteristics in a given season, including age and talent. Importantly, since protectors as defined in this study are constant throughout a season, these fixed effects also control for any characteristics of the batter's protector(s), including his approximate age. This is important because the protector's age could plausibly be correlated with both the timing of the protector's injury and the protectee's performance. As an aside, I should note that equation (5) is equivalent to a comparison of batting outcomes for a given batter in a given year broken down by whether his protector is injured

at the beginning of the season few players are injured and cold weather discourages extra base hits.

DOI: 10.2202/1559-0410.1296

or not. In other words, the reduced form takes the method originally used by James (1985) to study protection and applies it to all batters in MLB over several years.

Equation (5) is useful because it focuses on the protector injury variable generated by a quasi-random experiment. However, one would like to estimate how much hitting outcomes respond to a change in protector quality. We are really interested in an equation analogous to (2). Instrumental variables (IV) analysis does exactly this. Essentially, IV boils down to a two-stage process. The first stage is a regression of the endogenous variable (OPS of the actual on-deck hitter) on the instrument (injury to the batter's protector(s)) and controls, i.e. a regression of equation (4). The second stage follows the main equation (2) but differs in that, instead of using the OPS of the on-deck hitter, it uses the predicted values from the first stage and the standard errors are adjusted.[10]

2.4 Non-Independent Events

The above discussion directs us toward a way to get a consistent point estimate of protection on batting outcomes, whether that be by reduced form or IV, using the natural experiment caused by injuries to on-deck hitters. However, we of course need a confidence interval for these estimates as well. Frequently, researchers estimate standard errors that rely on the assumption that the error terms are independent across all observations. This is unlikely to be true for play-by-play baseball data. Consider the original Bill James study of Dale Murphy's performance and Bob Horner's injuries. During that stretch, Murphy actually performed slightly better when Horner was hurt, and with conventional standard errors, a large negative impact of Horner's injuries on Murphy's performance can be rejected. However, in the 1984 season, Horner played only 32 games, all of which were played in April or the second half of May. He sat the rest of the season with injuries. To make a statement about protection, we're comparing these games from April and half of May with the rest of the season. Even in the unlikely event that the month of the season would be uncorrelated with batting outcomes over a large sample of plate appearances, it's possible that Atlanta was hit by streak of bad weather that April or that Murphy was 'cold' or unlucky to start that year or that the Braves faced a string of umpires with small strike zones in the latter half of May. We expect these effects (except possibly the month) to be unrelated to injuries over

[10]See Angrist and Pischke (2009) for a guide to IV analysis as well as the experimental approach to regression.

the course of many seasons and players. Thus, we legitimately omit them from the regression. However, these random factors affect groups of consecutive plate appearances, introducing more uncertainty than what is captured by conventional standard errors. Any differences (or lack of differences) between batting lines in two groups of largely consecutive games can be explained by persistent random events. In effect, the sample size isn't really as large as we think because non-independent events give us less information than independent events. Our standard errors need to reflect this.

More generally, an important characteristic of (observable) injuries is that they almost always remove players from a large number of consecutive games. Batting outcomes are likely positively correlated within a game or over consecutive games due to weather, umpire tendencies, etc. If any of these are excluded from the regression, then they are in the error term. As a result, the error terms are very unlikely to be independent across a particular player's plate appearances and conventional standard errors will understate uncertainty. To correct for this, I will use cluster-robust standard errors, clustered at the batter-year level. For samples with a large number of batter-years, these estimates of the standard errors adjust the typical standard errors to allow for any possible correlation structure between outcomes of plate appearances within a given batter-year.[11] This allows for random events that affect multiple games in a given season.

3 Data

3.1 Data and Variable Definitions

Data for this study come from three sources. Play-by-play data on the batting outcome of every plate appearance in MLB for 2002-2009 was obtained free of charge from and is copyrighted by Retrosheet.[12] I exclude events that do not end the plate appearance (stolen bases, etc.) and plate appearances for which I don't observe the next batter (end of the game). Batting outcomes are drawn from this data and coded as dummy variables. I use a dummy for the general outcome of reaching base by walk, hit or being hit by a pitch. Hits are broken down further into singles, doubles, triples, and homeruns.

[11] These can be calculated using standard statistical software, including the ivreg2 package in STATA. For the theory see Moulton (1990) and for some theory, practical application in economics, and formulas for clustered standard errors, see Duflo, et al (2003) and chapter 8 of Angrist and Pischke (2009).

[12] Interested parties may contact Retrosheet at 20 Sunset Rd. Newark, DE 19711 or go to www.retrosheet.org.

DOI: 10.2202/1559-0410.1296

Table 1: Summary Statistics

	All Batters			Only Third Batters		
Variable	Observations	Mean	Std. Dev.	Observations	Mean	Std. Dev.
Total Bases	1500301	0.37	0.83	171519	0.43	0.92
On Base	1500301	0.33	0.47	171519	0.37	0.48
On Base + Total Bases	1500301	0.71	1.19	171519	0.80	1.27
Extra-base Hit	1500301	0.08	0.27	171519	0.10	0.30
Hit	1500301	0.24	0.42	171519	0.25	0.43
Single	1500301	0.16	0.36	171519	0.16	0.36
Double	1500301	0.05	0.21	171519	0.05	0.23
Triple	1500301	0.005	0.07	171519	0.004	0.06
Homerun	1500301	0.03	0.16	171519	0.04	0.19
Walk	1500301	0.09	0.28	171519	0.11	0.31
Unintentional Walk	1500301	0.08	0.27	171519	0.10	0.30
Intentional Walk	1500301	0.007	0.08	171519	0.012	0.11
Hit By Pitch	1500301	0.009	0.10	171519	0.010	0.10
Reached On Error	1500301	0.009	0.10	171519	0.008	0.09
Fielder's Choice	1500301	0.003	0.05	171519	0.002	0.05
Strikeout	1500301	0.17	0.38	171519	0.15	0.36
General Out	1500301	0.49	0.50	171519	0.46	0.50
On-deck OPS	1461356	0.76	0.14	167751	0.83	0.11
On-deck Injured	1500301	0.03	0.18	171519	0.08	0.28
NL	1500301	0.53	0.50	171519	0.52	0.50
Platoon	1500301	0.54	0.50	171519	0.52	0.50
ERA	1478090	4.50	2.30	168952	4.52	2.47

Of these, all but singles are classified as extra-base hits. Walks are broken down into intentional walks and non-intentional walks. Batters hit by pitches are measured separately from walks. I also use a total bases outcome, which similar to slugging, assigns to each hit outcome the number of bases attained on the hit (i.e. 3 for a triple).[13] I also include indicators for reaching base on an error, fielder's choice, striking out, and other outs. Ballpark, out, inning, National League, runner, and month dummies are coded from this data as well. Platoon advantage of the batter, defined as when the batter and pitcher are opposite-handed,[14] is also coded from this data.

The onbase plus slugging (OPS) of the batter that actually follows the current batter is calculated using this data. OPS is a combination of two statistics: onbase percentage and slugging average. Onbase percentage (OBP) is defined as the percentage of plate appearances that a batter reaches base:

$$OBP = \frac{Hits+NonIntentionalWalks+Hitbypitch+IntentionalWalks}{PlateAppearances}$$

Slugging (SLG) is defined as total bases divided by at-bats and gives a sense of how much power a player hits with:

[13]In the case of some fielding errors, there is a difference between the number of bases reached by the batter and the number credited to the hit. I use those credited to the hit.

[14]In baseball, being opposite-handed gives the batter a noticeable advantage.

$$SLG = \frac{Singles+2*Doubles+3*Triples+4*Homeruns}{AtBats}$$

These are both calculated using Retrosheet data for each batter-year. Finally, OPS is the sum of OBP and SLG, and OPS of the on-deck hitter is assigned using Retrosheet data on which player follows the current batter. While somewhat less intuitive to interpret than its components, OPS is one decent way of combining a measure of getting on base and a measure of power hitting. Though many baseball statisticians prefer other measures, OPS has the value of simplicity, ability to predict future performance, common usage, and high correlation with more advanced measures. Thus, I use it as a measure of the on-deck batter's overall offensive ability. Also, for individual plate appearances, I will use an analogous outcome variable to measure overall offensive output: the sum of total bases and the on-base indicator. This measure is not exactly the same as OPS since slugging uses only at-bats, not all plate appearances, but I will interpret it as a rough measure of OPS for a particular plate appearance.

The second major source of data for this study is a database of all disabled list trips for all players in MLB 2002-2009 created by Josh Hermsmeyer (blog.rotobase.com). Given the data available, I only consider injuries that result in a trip to the disabled list and thus define a player as injured if they are on the disabled list. To denote particular plate appearances as being affected by a protector injury, I need to define a notion of protection that is not affected by the injuries themselves. This is a surprisingly difficult task because we want to know which player would have protected a given batter if there were no injuries. Unfortunately, we cannot observe this precisely when we need to, when injuries do in fact occur.

As a result, there are many possible methods for designating particular plate appearances as affected by a protector injury. For clarity, I intentionally choose a very broad measure and then narrow it explicitly. Consider a player A in year Y. I use the Retrosheet play-by-play data to determine which player, B, preceded batter A in the batting order most frequently in year Y. Call A a protector of B. Defined in this way, a player may have multiple protectors because protection is defined relative to the protector's most frequent protectee, not the protectee's most frequent protector.[15] The broadest measure of an injured protector, then, is an indicator of whether any of the batter's protectors are injured during the plate appearance. However, the broadest definition is

[15]In an earlier version, I used the latter definition, and the results are very similar and available upon request. I discard this earlier definition because it has the unfortunate property that an injury to a good player that lasts more than half of the season can result in them not being assigned as a protector of anyone.

DOI: 10.2202/1559-0410.1296

a bit too broad. In this instance, many players are 'protected' by poor players including a large number of pitchers and part-time minor leaguers. Not surprisingly, injuries to bullpen pitchers and bench-warmers do not generate a decrease in the quality of the on-deck hitter, i.e. these injuries provide poor natural experiments for studying protection.

To focus the analysis on players who are actually providing protection, I will only consider players to be protectors if they have more than 200 plate appearances (about one third of a season) and have an OPS above .750 (about league average) in the given season. Limiting the pool of protectors in this way guarantees that injuries to these players do in fact cause a decrease in the quality of the on-deck hitter, while still including a broad range of every-day players as well as those who are injured for long periods of time. Other than this limitation in who can be counted as a protector, I will use the definition of the previous paragraph. Thus, players with more than 200 PA and .750 OPS will be protectors of the players that most frequently precede them in the batting order, and a particular plate appearance will be considered subject to an on-deck injury if any of that player's protectors are hurt at the time.

Finally, I take the ERA control variable from the Lahman Baseball Database (Lahman 2009). This variable is assigned to plate appearances based on the pitcher listed by Retrosheet as pitching during that plate appearance. ERA denotes 'earned run average' and is a commonly used measure of pitcher performance. It measures the number of earned (as opposed to resulting from fielding errors) runs allowed by a pitcher per nine innings. While not always favored by statisticians, it has the advantage of easy interpretation, common use, and fair performance prediction. Also, I prefer ERA to explicitly using the hit, walk, homerun, etc. rates of the pitcher due to bias created by using controls that are themselves averages of the dependent variable.[16]

3.2 Summary Statistics

The first pane of Table 1 provides summary statistics for the whole sample. Batters get hits about 24 percent of the time and walk 9 percent of the time; thus, they get on base (not including errors and fielder's choice) about 33 percent of all plate appearances. Extra-base hits occur about 8 percent of the time, and batters average about .37 total bases from hits per plate appearance. Most other events are rather rare, except strikeouts (17 percent) and other outs (49 percent). The actual hitters that bat after the current batters have an average OPS of 0.755. Meanwhile, batters are hit by a protector injury

[16]See Angrist and Pischke (2009)

Table 2: First-Stage

Dependent Variable: On-deck OPS

Variable	All Batters	1	2	3	4	5	6	7	8	9
Ondeck Injured	-0.028***	-0.027***	-0.033***	-0.029***	-0.033***	-0.021**	-0.015	-0.028**	0.015	-0.057***
	(0.004)	(0.006)	(0.011)	(0.008)	(0.008)	(0.009)	(0.010)	(0.014)	(0.031)	(0.015)
NL	-0.022*	-0.009	-0.020	0.010	0.037	-0.000	0.027	-0.080***	-0.109***	-0.079
	(0.013)	(0.027)	(0.022)	(0.033)	(0.038)	(0.021)	(0.081)	(0.014)	(0.021)	(0.059)
Platoon	0.006***	-0.002	0.005***	0.002	0.001	0.008***	0.012***	0.014***	0.013***	0.000
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
ERA	0.00009	0.00049***	0.00024	0.00020*	-0.00010	-0.00015	0.00005	0.00006	-0.00064**	0.00024
	(0.00007)	(0.00015)	(0.00018)	(0.00010)	(0.00015)	(0.00017)	(0.00018)	(0.00019)	(0.00029)	(0.00018)
Batter-Year FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Runner Dummies	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Out Dummies	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Inning Dummies	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Ballpark Dummies	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Month Dummies	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES

Statistical significance at the 1, 5, and 10 percent levels is denoted by ***, **, and * respectively.
Standard errors are clustered at the batter-year level.

about 3 percent of the time. Plate appearances are slightly more likely than not to be in the National League and to have a batter with platoon advantage. Finally, the ERA of opposing pitchers averages at about 4.5 earned runs per nine innings.

The second pane of Table 1 provides the same statistics for players who most frequently bat third in the batting order. As expected, they are better than average hitters, performing better across all batting outcomes (except triples). They get on base more often and hit with greater power. For example, their homerun rate is 3.9 percent which is about 1.4 times the rate of an average player. They also tend to have better batters following them. Their on-deck hitters have OPS values that are about 75 points higher than the average on-deck hitter, but their usual protectors are hurt 8 percent of the time. The higher than average injury rate mainly reflects the fact that the batters following them qualify for the protector cutoffs of 200 PA and .750 OPS more often.

4 Results

4.1 Relevance

First, we need to confirm that an injury to the batter's protector(s) really does lower the quality of the batter that actually follows him. For this, we want to estimate equation (4) by ordinary linear regression. The first column of Table 2 provides the results of this estimation over the entire sample. The coefficient on injuries is negative, statistically significant at the 1 percent level, and meaningful in size. The coefficient of -0.028 indicates that when a batter's

DOI: 10.2202/1559-0410.1296

protection goes to the disabled list, the batter gets an actual on-deck hitter with an OPS that is, on average, 28 points lower, or about one fourth of a standard deviation.[17] This difference represents a meaningful drop in the average quality of protection.

The remaining columns of Table 2 slice the sample according to the batter's usual lineup position.[18] The results are as expected. Most teams place their best batters in the top half of the order, so injuries have greatest impact on the protection of these players (as well as 9th hitters who are followed by the first hitter). Injuries to the protectors of the first through fifth hitters cause drops in protection ranging from 21 to 33 points of OPS. These effects are less apparent in the bottom half of the batting order, both in magnitude and statistical significance, and the point estimate is even positive for eighth hitters.

Aside from averages, the distribution of changes in protection becomes important when assessing the applicability of this study. If all changes in protection induced by injuries are modest, then it would be invalid to apply the results of this study to large changes in protection. For batter-years that suffer protector injuries, I calculate the difference in average on-deck OPS with the protector healthy as opposed to injured. Figure 1 displays the distribution of these injury-induced changes in protection over the whole sample. As expected from the above results, the mean appears to be negative, though some injuries do result in better protection (positive values). Importantly, while the mean change in protection caused by injury is only 28 points, there are still a large number of cases with very large changes in protection, around 100 to 200 points. Figure 2 provides the same distribution for third hitters only. Again, injuries most frequently hurt protection and the distribution includes large swings in protection, with many batters having their protection drop by 100 points of OPS or more.[19]

[17]Throughout I use 'points' to denote the thousandths place. For percentages, this corresponds to tenths of a percentage point. The standard deviation of OPS is calculated over the population of players with at least 200 plate appearances, not over at bats. Hence, it differs from that reported in Table 1.

[18]Usual is defined as most-frequent lineup position for that batter in the given year.

[19]Although this paper focuses on injuries, the coefficients on NL in Table 2 follow the expected patterns. Batters receive less protection in the National League, and this effect comes almost entirely through seventh and eighth batters, not surprising given that eighth hitters are followed by pitchers in the NL.

Figure 1: Distribution of Injury-Induced Protection Changes, All Hitters

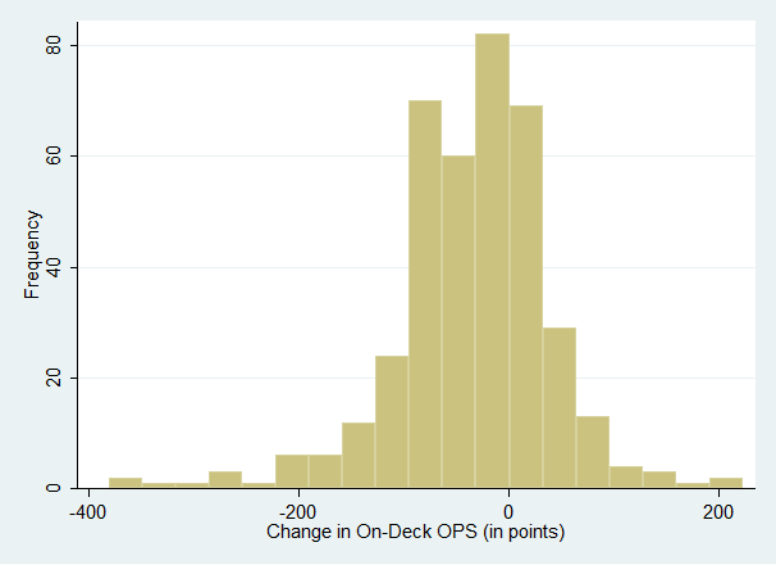
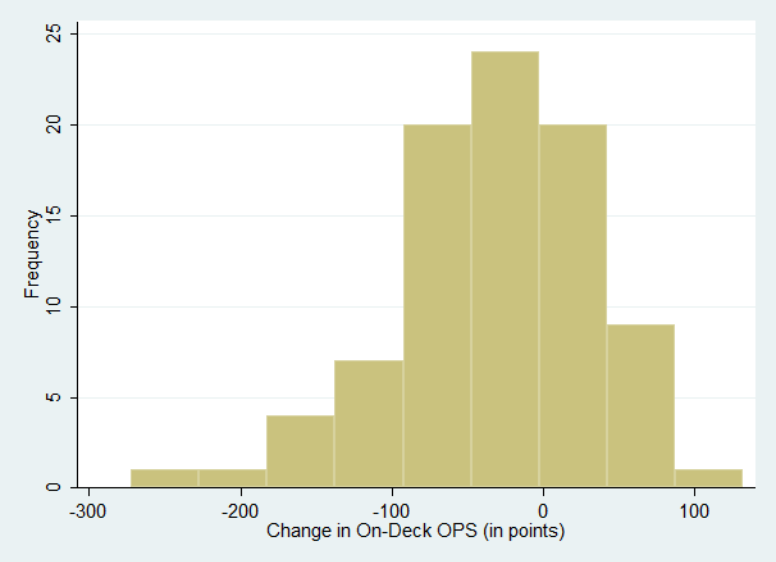


Figure 2: Distribution of Injury-Induced Protection Changes, Third Hitters



4.2 Reduced Form

We have confirmed the fact that injuries to a batter's protector(s) do have a significant negative impact on the quality of protection. If we believe that, conditional on the controls, this is the only path through which these injuries affect batter outcomes, we can proceed treating these injuries as a natural experiment that lowers the quality of protection. I start by estimating the reduced form, equation (5), regressing the various batting outcomes on the protector injury dummy and control variables taking into account batter-year fixed effects. The results for all hitters are displayed in Table 3. Statistically, the strongest result relates to intentional walks which are significant at the 1 percent level. Batters left unprotected by injury are intentionally walked .18 percentage points more often. This may seem small, but relative to a mean likelihood of 0.7 percent across the sample, this represents a large increase in the probability of being intentionally walked. This result, though, is simply an unsurprising and non-controversial confirmation of the 'weak' form of protection. Protection exists in the very specific, uncommon situations in which intentional walks are used.[20]

For extra-base hits, the coefficient on protector injuries is negative over the whole sample, but misses statistical significance. The point estimate indicates that the probability that a batter hits an extra-base hit falls .23 percentage points if his protection is injured. Given that batters only hit extra-base hits about 8 percent of the time, this represents a 2.9 percent drop in extra-base hits, relative to the mean. This does not appear to be simply due to the fact that the batters put fewer balls into play. The likelihood of a single rises slightly. These are significant changes in baseball terms, especially when one considers that these are responses to a decrease in protection quality of only one-quarter of a standard deviation. If the effects of protection are roughly linear, a one standard deviation drop in protection could decrease extra-base hits by about 12 percent.

The effect of protection for the whole sample could be large but is imprecisely measured due to the rarity of the injuries and the use of clustered standard errors. This led to estimates that just miss statistical significance. To further investigate whether a real protection effect is driving these results, we should focus on where we expect to observe protection: hitters batting third in the lineup. Commentators and baseball practitioners most frequently invoke the theory of protection when describing the performance of the very best hitters and discussing the need to protect the number 3 with a good cleanup

[20]Generally in close games with first base open and at least one out. See Tango, et al (2007) for a detailed analysis of these situations.

hitter. Additionally, as we saw above, injuries to on-deck hitters affect number 3 hitters the most.

Table 4 replicates the reduced form results but only for batters that most frequently bat in the third spot in the batting order. The results are strikingly similar to what the protection hypothesis would predict. Weak protection is easily confirmed as third hitters are walked, both intentionally and overall, much more when their protection is injured. The result for intentional walks is about 2 times larger than it was for all hitters. The probability of a hit doesn't change, but the distribution gets skewed toward singles and away from homeruns and doubles, with the latter two significant at the 10-percent level. Extra-base hits drop by 0.63 percentage points when the batter is struck by a protector injury. These all indicate a very strong effect of protection on the *distribution* of batting outcomes for number three hitter.

Unfortunately, I am unable to make statistical statements about the impact on overall production. More walks and fewer extra-base hits lead to statistically insignificant increases in getting on-base and decreases in total bases. If I measure total production by their sum (an analogue of OPS), then being left unprotected leads to a statistically insignificant 6 point drop.[21]

4.3 Instrumental Variables

Having detected an impact of protection above, it would be useful to quantify the impact of protection in terms of the change in probability of a particular batting outcome in response to changes in a measure of the quality of protection (OPS). This has the advantage of being more useful to baseball practitioners and more comparable to previous results in the literature. I start by confirming some results of the previous literature for this sample. First, I estimate a simple ordinary least squares regression (linear probability model) of batting outcomes on the OPS of the on-deck hitter, as specified in equation (1). The first column of Table 5 reports the coefficient on on-deck OPS for each batting outcome. Not surprisingly, this specification results in statistically significant, fairly large, and almost uniformly positive coefficients for positive outcomes. The one exception is the strong, negative coefficient on intentional walks, showing that the weak form of protection can be found even in the simple correlations. Of course, these results are not credible. They mostly reflect the fact that good hitters are clumped together in the typical

[21]Third hitters are the batters most strongly affected by protection. Other batters from 1st through 4th exhibit similar patterns in the point estimates but without statistical significance.

DOI: 10.2202/1559-0410.1296

lineup, but these are the effects we would measure if we ignored all confounding variables. The second column of Table 5 adds in batter-year fixed effects to control for the ability of batters as well as any other batter characteristics that are constant across a given season. This changes the picture entirely, producing mostly statistically significant, almost uniformly negative (for positive outcomes), and relatively small coefficients on the quality of protection. Adding all of the other control variables in the third column does not change these results greatly. These results are qualitatively similar to previous results produced by Bradbury and Drinen (2008) using similar specifications but covering a different time period. They interpret this as evidence that either protection does not exist or that pitchers exert greater effort during particularly important situations (i.e. when good batters are protected) so that protection, at least in practice, has no noticeable effect.

As discussed above, selective choice of the batting order by managers based on expected performance could create this negative correlation as well. So, we turn to the natural experiment caused by injuries. The final column of Table 5 displays the estimates from an IV specification where injuries to protectors are used as the instrument for the OPS of the actual on-deck hitter. As usually occurs, the IV results closely follow the reduced form results. The results indicate that when backed by better protection, batters hit more extra-base hits, which is now significant at the 10 percent level. Additionally, they are walked less, both intentionally and in general. For interpretation, the coefficient of 0.094 on extra-base hits indicates that a 100 point (about one standard deviation) increase in the OPS of the on-deck hitter results in a .94 percentage point increase in the probability of an extra-base hit.

Table 6 replicates Table 5 for third hitters only. The simple OLS reflects that better star hitters are on better teams and thus have better hitters following them. Again, intentional walks are the exception. Adding fixed effects and controls results in small and statistically insignificant coefficients, except for walks. Moving to the IV once again results in a strong verification of the theory of protection. The coefficients are much larger than for the entire sample, as expected, and the pattern of signs and statistical significance reflects a shift toward extra-base hits and away from singles and walks when batters have better protection behind them. Also, batters strike out more often as pitchers throw more strikes. Clearly, the distribution of batting outcomes reacts to protection for these elite hitters.

Table 3: Reduced Form, **All Batters**

	Ondeck Injured	NL	Platoon	ERA
Hit	-0.0018	0.0297	0.0120***	0.0066***
	(0.0024)	(0.0182)	(0.0009)	(0.0003)
Total Bases	-0.0066	0.0775**	0.0354***	0.0142***
	(0.0047)	(0.0310)	(0.0018)	(0.0006)
Onbase	0.0016	0.0231	0.0326***	0.0097***
	(0.0027)	(0.0176)	(0.0010)	(0.0004)
XBHit	-0.0023	0.0289**	0.0118***	0.0039***
	(0.0015)	(0.0118)	(0.0006)	(0.0002)
Onbase + Total Bases	-0.0050	0.101**	0.0680***	0.0239***
	(0.0066)	(0.0433)	(0.0026)	(0.0010)
Single	0.0006	0.000839	0.0002	0.0027***
	(0.0022)	(0.0198)	(0.0008)	(0.0002)
Double	-0.0011	0.0169*	0.0059***	0.0019***
	(0.0012)	(0.00967)	(0.0004)	(0.0001)
Triple	0.0001	0.00495	0.0003**	0.0003***
	(0.0004)	(0.00447)	(0.0001)	(0.0000)
Homerun	-0.0013	0.00700	0.0056***	0.0018***
	(0.0009)	(0.00728)	(0.0004)	(0.0001)
Walk	0.0029	-0.0115	0.0261***	0.0028***
	(0.0018)	(0.0143)	(0.0007)	(0.0002)
Non-IBB	0.0012	-0.0111	0.0172***	0.0028***
	(0.0017)	(0.0145)	(0.0006)	(0.0002)
IBB	0.0018***	-0.000426	0.0089***	0.0001**
	(0.0006)	(0.00264)	(0.0003)	(0.0000)
HBP	0.0004	0.00499	-0.0056***	0.0003***
	(0.0005)	(0.00456)	(0.0002)	(0.0000)
Reached On Error	-0.0001	0.00477	-0.0007***	0.0000
	(0.0005)	(0.00679)	(0.0002)	(0.0000)
Fielder's Choice	0.0001	0.000191	-0.0008***	-0.00004***
	(0.0002)	(0.00358)	(0.0001)	(0.00001)
Strikeout	-0.0025	-0.0272	-0.0264***	-0.0060***
	(0.0021)	(0.0201)	(0.0009)	(0.0002)
General Out	0.0009	-0.000882	-0.0047***	-0.0036***
	(0.0028)	(0.0244)	(0.0011)	(0.0002)

Statistical significance at the 1, 5, and 10 percent levels is denoted by ***, **, and * respectively. All regressions use batter-year fixed effects and control for out, inning, runner, ballpark, and month dummies. Standard errors are clustered at the batter-year level.

DOI: 10.2202/1559-0410.1296

Table 4: Reduced Form, **3rd Batters**

	Ondeck Injured	NL	Platoon	ERA
Hit	-0.0011	-0.0115	0.0126***	0.0057***
	(0.0048)	(0.0419)	(0.0028)	(0.0008)
Total Bases	-0.0135	0.0700	0.0453***	0.0131***
	(0.0100)	(0.0994)	(0.0060)	(0.0018)
Onbase	0.0071	-0.0245	0.0448***	0.0088***
	(0.0059)	(0.0218)	(0.0032)	(0.0011)
XBHit	-0.0063*	0.0708**	0.0166***	0.0037***
	(0.0032)	(0.0320)	(0.0019)	(0.0006)
Onbase + Total Bases	-0.0064	0.0455	0.0901***	0.0219***
	(0.0143)	(0.1200)	(0.0084)	(0.0029)
Single	0.0052	-0.0823*	-0.0039	0.0020***
	(0.0037)	(0.0425)	(0.0024)	(0.0004)
Double	-0.0038*	0.0648***	0.0083***	0.0018***
	(0.0023)	(0.0010)	(0.0013)	(0.0003)
Triple	0.0009	0.0013	0.0005	0.0001
	(0.0008)	(0.0012)	(0.0004)	(0.0000)
Homerun	-0.0035*	0.0047	0.0078***	0.0019***
	(0.0018)	(0.0240)	(0.0013)	(0.0003)
Walk	0.0081*	-0.0180	0.0389***	0.0029***
	(0.0044)	(0.0252)	(0.0021)	(0.0005)
Non-IBB	0.0038	0.0187	0.0225***	0.0027***
	(0.0040)	(0.0288)	(0.0019)	(0.0005)
IBB	0.0044***	-0.0367***	0.0164***	0.0002
	(0.0015)	(0.0090)	(0.0009)	(0.0001)
HBP	0.0001	0.0050	-0.0068***	0.0002**
	(0.0010)	(0.0066)	(0.0007)	(0.0001)
Reached On Error	-0.0017**	0.0371*	-0.0018***	-0.0001
	(0.0008)	(0.0209)	(0.0005)	(0.0001)
Fielder's Choice	-0.0002	0.0270	-0.0009***	0.0000
	(0.0005)	(0.0237)	(0.0003)	(0.0000)
Strikeout	-0.0086**	-0.0782***	-0.0324***	-0.0046***
	(0.0040)	(0.0168)	(0.0023)	(0.0006)
General Out	0.0033	0.0376	-0.0098***	-0.0040***
	(0.0059)	(0.0473)	(0.0032)	(0.0007)

Statistical significance at the 1, 5, and 10 percent levels is denoted by ***, **, and * respectively. All regressions use batter-year fixed effects and control for out, inning, runner, ballpark, and month dummies. Standard errors are clustered at the batter-year level.

Table 5: OLS and Instrumental Variables, **All Batters**

	OLS	OLS - FE	OLS - FE and Controls	IV
Hit	0.037***	-0.011***	-0.009***	0.078
	(0.003)	(0.003)	(0.003)	(0.090)
Total Bases	0.102***	-0.021***	-0.016***	0.279
	(0.006)	(0.006)	(0.006)	(0.174)
Onbase	0.050***	-0.038***	-0.041***	-0.048
	(0.004)	(0.003)	(0.004)	(0.096)
XBHit	0.027***	-0.006***	-0.004**	0.094*
	(0.002)	(0.002)	(0.002)	(0.057)
Onbase + Total Bases	0.152***	-0.059***	-0.057***	0.230
	(0.009)	(0.009)	(0.009)	(0.242)
Single	0.010***	-0.005**	-0.005*	-0.015
	(0.003)	(0.003)	(0.003)	(0.079)
Double	0.008***	-0.004**	-0.003*	0.042
	(0.001)	(0.002)	(0.002)	(0.044)
Triple	0.000	0.000	0.001	-0.003
	(0.000)	(0.001)	(0.001)	(0.016)
Homerun	0.019***	-0.002**	-0.002*	0.055
	(0.001)	(0.001)	(0.001)	(0.034)
Walk	0.013***	-0.028***	-0.032***	-0.111*
	(0.003)	(0.002)	(0.002)	(0.065)
Non-IBB	0.033***	0.002	-0.001	-0.044
	(0.003)	(0.002)	(0.002)	(0.062)
IBB	-0.019***	-0.029***	-0.031***	-0.066***
	(0.001)	(0.001)	(0.001)	(0.022)
HBP	-0.000	0.000	0.000	-0.016
	(0.000)	(0.001)	(0.001)	(0.019)
Reached on Error	0.001	0.002**	0.002***	0.005
	(0.001)	(0.001)	(0.001)	(0.018)
Fielder's Choice	-0.001	0.000	0.000	-0.003
	(0.000)	(0.000)	(0.000)	(0.009)
Strikeout	-0.031***	0.012***	0.008***	0.091
	(0.004)	(0.003)	(0.003)	(0.075)
General Out	-0.019***	0.024***	0.031***	-0.045
	(0.005)	(0.004)	(0.004)	(0.099)

Statistical significance at the 1, 5, and 10 percent levels is denoted by ***, **, and * respectively. Coefficients on on-deck OPS reported. Standard errors are clustered at the batter-year level.

DOI: 10.2202/1559-0410.1296

Table 6: OLS and Instrumental Variables, **3rd Batters**

	OLS	OLS - FE	OLS - FE and Controls	IV
Hit	0.021*	0.007	0.013	0.075
	(0.011)	(0.010)	(0.011)	(0.173)
Total Bases	0.091***	0.026	0.033	0.571
	(0.024)	(0.022)	(0.023)	(0.388)
Onbase	0.029**	-0.027**	-0.017	-0.209
	(0.012)	(0.011)	(0.011)	(0.201)
XBHit	0.025***	0.010	0.012	0.254**
	(0.007)	(0.007)	(0.007)	(0.128)
Onbase + Total Bases	0.120***	-0.001	0.016	0.362
	(0.031)	(0.029)	(0.030)	(0.522)
Single	-0.004	-0.004	0.001	-0.179
	(0.010)	(0.009)	(0.009)	(0.132)
Double	0.003	0.006	0.008	0.149*
	(0.006)	(0.006)	(0.006)	(0.087)
Triple	-0.000	-0.001	-0.001	-0.033
	(0.002)	(0.002)	(0.002)	(0.031)
Homerun	0.023***	0.005	0.004	0.137*
	(0.006)	(0.005)	(0.005)	(0.073)
Walk	0.008	-0.037***	-0.030***	-0.282*
	(0.015)	(0.010)	(0.009)	(0.159)
Non-IBB	0.029**	0.002	0.000	-0.130
	(0.014)	(0.008)	(0.001)	(0.142)
IBB	-0.026***	-0.037	-0.031***	-0.152***
	(0.004)	(0.005)	(0.004)	(0.057)
HBP	0.005	0.001	0.001	-0.003
	(0.003)	(0.002)	(0.002)	(0.036)
Reached on Error	-0.000	0.001	0.002	0.060*
	(0.009)	(0.002)	(0.003)	(0.033)
Fielder's Choice	0.003***	0.002*	0.002	0.007
	(0.001)	(0.001)	(0.001)	(0.017)
Strikeout	-0.018	-0.001	-0.001	0.351**
	(0.017)	(0.010	(0.010)	(0.176)
General Out	-0.014	0.025**	0.014	-0.207
	(0.019)	(0.012)	(0.013)	(0.227)

Statistical significance at the 1, 5, and 10 percent levels is denoted by ***, **, and * respectively. Coefficients on on-deck OPS reported. Standard errors are clustered at the batter-year level.

The interpretation of the coefficients for total bases and getting on base are perhaps most interesting, though they are statistically insignificant. The coefficient of 0.571 on total bases indicates that for a 100 point increase in on-deck OPS, the batter reaches about .06 more bases. This means that, roughly, for every point that the on-deck hitter's OPS increases, the slugging of the batter goes up by about 0.6 points.[22] Similarly, the on-base percentage of the batter decreases by about .21 points when he has a protector with an OPS that is one point higher. These are very large but off-setting effects, in terms of overall production. The point estimate on the sum of on-base and total bases is positive, but statistically insignificant. As a result, though the data indicate that protection does affect the distribution of batting outcomes, I cannot make a statement about the impact of protection on overall production.

To get a sense of the impact of a large change in protection, I compute the effect of a 100 point increase in the OPS of the on-deck hitter and then compare this to the mean outcome. Table 7 displays these results, for all hitters and for number three hitters. For example, all hitters have a mean probability of hitting an extra-base hit of 8 percentage points. The 0.94 percentage point increase in the probability of an extra-base hit induced by better protection means that batters are hitting 9.67 percent more extra-base hits than average. For third hitters, the effect is much larger, resulting in a 26.25 percent increase in extra-base hits, even though third hitters have a higher extra-base hit rate to begin with. Clearly, these are effects large enough to change the way baseball is played and managed.

For comparison, the same numbers are listed for the specification using ordinary linear regression, batter-year fixed effects, and the common list of control variables. This helps contrast the ordinary regression approach and the natural experiment approach taken in this paper. Where IV estimates indicate an increase of 9.7 percent, ordinary regressions suggest that a 100 point increase in protector OPS would decrease extra-base hits by about a half of a percent. For third hitters, the contrast is even starker. Ordinary regressions almost always suggest very small changes in batting outcomes, but IV estimates based on player injuries indicate that 26 percent more extra-base hits, 11 percent fewer singles, 28 percent more doubles, 35 percent more homeruns, and 26 percent fewer walks result from such an improvement in protection. Only regarding intentional walks do the two methods agree on the direction and importance of the impact of protection, but even here, the results based on the natural experiment demonstrate larger effects of protection.

[22]This is roughly, not exactly, because slugging is total bases per at bat while I am dealing with total bases per plate appearance. At bats exclude some events that plate appearances include, notably walks. So, slugging would actually rise by somewhat more than 0.6 points.

DOI: 10.2202/1559-0410.1296

Table 7: Estimated Effects Relative to Means

Method Change in On-Deck OPS	Mean	OLS 100 Points	IV 100 Points	IV 20 Points
All Batters				
Hit	0.24	-0.38%	3.32%	0.66%
Total Bases	0.37	-0.43%	7.44%	1.49%
Onbase	0.33	-1.24%	-1.46%	-0.29%
XBHit	0.08	-0.41%	9.67%	1.93%
Onbase + Total Bases	0.71	-0.70%	2.86%	0.57%
Single	0.16	-0.32%	-0.98%	-0.20%
Double	0.05	-0.63%	8.83%	1.77%
Triple	0.005	2.04%	-6.78%	-1.36%
Homerun	0.03	-0.73%	20.13%	4.03%
Walk	0.09	-3.74%	-12.97%	-2.59%
NIBB	0.08	-0.13%	-5.72%	-1.14%
IBB	0.007	-43.93%	-93.25%	-18.65%
HBP	0.009	0.00%	-17.16%	-3.43%
Reached on Error	0.009	2.10%	5.87%	1.17%
Fielder's Choice	0.003	1.81%	-11.15%	-2.23%
Strikeout	0.17	0.44%	5.35%	1.07%
General Out	0.49	0.63%	-0.93%	-0.19%
Third Batters				
Hit	0.25	0.51%	2.97%	0.59%
Total Bases	0.43	0.76%	13.22%	2.64%
Onbase	0.37	-0.46%	-5.60%	-1.12%
XBHit	0.10	1.24%	26.25%	5.25%
Onbase + Total Bases	0.80	0.19%	4.50%	0.90%
Single	0.16	0.06%	-11.44%	-2.29%
Double	0.05	1.49%	27.67%	5.53%
Triple	0.004	-2.49%	-80.79%	-16.16%
Homerun	0.04	1.03%	35.22%	7.04%
Walk	0.12	-2.72%	-25.60%	-5.12%
NIBB	0.10	0.00%	-13.29%	-2.66%
IBB	0.012	-25.18%	-123.44%	-24.69%
HBP	0.010	1.05%	-2.79%	-0.56%
Reached on Error	0.008	2.34%	70.97%	14.19%
Fielder's Choice	0.002	8.58%	29.84%	5.97%
Strikeout	0.15	-0.05%	22.80%	4.56%
General Out	0.46	0.31%	-4.48%	-0.90%

When interpreting these results, it is important to consider the restriction imposed by a linear functional form. As noted above, the injuries studied here resulted in average drops in on-deck hitter quality of about 28 OPS points for the average batter and 29 OPS points for the average number three hitter. As shown above, the sample covers a wide distribution of changes in protection, so the results should be a good linear approximation of the impact of typical changes in protection. Still, if the impact of protection is non-linear, a 100 point change in OPS could have larger or smaller impacts than estimates extrapolated from this linear functional form. For example, the results for intentional walks seem unreasonably large, indicating that a 100 point increase in on-deck OPS will cut intentional walks by more than 100 percent. It seems likely here that the effect is not linear, with teams choosing to intentionally walk a batter if the gap in quality between the batter and the on-deck hitter is above some threshold and the situation allows for an intentional walk. Eventually, no matter how poor the hitter, advantageous intentional walk opportunities would no longer present themselves.[23] To provide another point of comparison, the final column of Table 7 provides the results for a 20 point OPS change for all hitters. Even so, this fairly minor change in the quality of protection still has a noticeable impact in shifting production from getting on base to hitting for power. All batters hit 2 percent more extra base hits than normally, and third batters hit 5 percent more. Together, these results confirm that large changes in protection could have dramatic impacts on batting outcomes and even modest changes in protection can have noticeable effects.

5 Conclusion

This study examined the theory that, in baseball, a batter's performance is affected by whether or not he is 'protected' by a high-quality hitter following him in the batting order. The strongest version of the theory predicts that hitters with better protection will be walked less (especially intentionally) while hitting for more power when they put the ball in play. A weaker version predicts only more walks. In analyzing play-by-play data from Major League Baseball for 2002-2009, I found evidence confirming the existence of the 'strong' form of protection, with batters not only being walked less but also having hits shifted toward extra-base hits. These effects were of a large

[23] In further analysis, available upon request, I have investigated non-linear effects and found some evidence that the impact on intentional walks does seem to get weaker for large drops in protection. However, the impact on extra-base hits actually appears to strengthen for large drops in protection.

DOI: 10.2202/1559-0410.1296

enough magnitude to affect batting outcomes in important ways. These effects are particularly strong among hitters placed third in the batting order. This strengthens the results because these are the batters for whom the protection hypothesis is often invoked in commentary on the game.

These results differ from previous studies of the protection hypothesis, which have unanimously found no or negligible evidence for protection. The main distinctive feature of this study is that I examine protection using random variation in on-deck batter quality caused by injuries. This approach differs from many previous studies that try to make comparisons within a particular batter-season or regress batting outcomes on on-deck hitter quality while controlling for as many important variables as possible. Both of these approaches regard on-deck batter quality as exogenously given, conditional on some observable variables. However, if on-deck hitter quality is largely the result of batting order choices made by a manager and if many of the factors affecting this choice both have a direct effect on batting outcomes and are unobservable, then estimates from these methods will be confounded by an endogeneity/omitted variable bias. This study attempts to overcome this source of bias by using the quasi-random variation in on-deck hitter quality created by injuries. This takes the literature back to the roots of the study of lineup protection embodied in Bill James' case study of Dale Murphy's performance and Bob Horner's injuries. Ironically, this expansion of James' original approach produces a different result.

On a technical side, this study also explicitly attempts to apply methods commonly used in other fields to study designed and natural experiments to the statistical study of baseball. In particular, I used clustered standard errors, which are extremely important to use in cases where the researcher recognizes the likelihood that several plate appearances (more generally observations) are affected by some common random effect that, while not creating bias in the point estimates, results in conventional standard errors underestimating the true level of uncertainty.

These results represent a first step toward testing whether protection has implications for how baseball is managed and researched. Contrary to previous statistical studies and in support of baseball's conventional wisdom, protection does appear to exist and exert a major influence on baseball at its highest levels, particularly among the best hitters. However, the overall effect of protection is ambiguous because the 'strong' form of protection detected in this study has two offsetting effects: more extra-base hits but fewer walks. When judging the effect on overall productivity, I find a positive, but statistically insignificant effect of protection on overall production, leaving us with no definitive answer on the overall impact of protection.

This ambiguous overall effect represents an important area of future research. First, the precision of the estimates could be improved. The rarity of trips to the disabled list and the need to cluster inevitably result in large standard errors. Even with the 1.5 million plate appearances in this dataset and an apparently large effect, the impact of protection is statistically significant for only a few batting outcomes. This could be improved by extending injury data to cover more seasons or to include all injuries, not just those that result in DL trips. This would serve to gather further evidence in a broader context and measure the effects more precisely. In particular, it would help measure the effect of protection on overall production to see if it is in fact positive.

Second, the protection theory would suggest that protection should be most important when the value of extra-base hits is high relative to the value of walks because it is in these situations that pitchers have an incentive to pitch around batters. Future research should focus on how the impact of protection differs across various runner, out, and inning game states. We should expect that protection has greatest force precisely when extra-base hits are most valuable. This can provide a test of the theory, and it would provide another outlet by which the impact of protection on overall production may be positive. Standard measures of overall production (like OPS) ignore the game state in which outcomes occur. If protection skews production towards the most valuable outcomes for a particular game state, then it would have a positive effect that these measures cannot capture but other measures can.

If protection is found to have a significant overall effect, then this will have implications for setting batting orders and valuing players. For example, teams may legitimately value a player both for his own production and the protection he provides for another player's production. However, most of these implications depend on whether the tendency toward more extra-base hits dominates the tendency toward fewer walks in terms of overall value. Thus, this study cannot provide a definitive answer on these questions.

Some implications can be gathered from the existing results, though. First, the existence of protection creates complications for research that attempts to identify optimal lineups using methods that assume independence of a batter's outcomes from the quality of the batter who is on-deck. At the very least, protection needs to be considered more carefully in such research. Second, the distributional impact of protection is itself important. Currently, common practice involves adjusting players' batting lines for their home ballpark when trying to compare different players. Given the impact of protection, players with poor protection will appear to be more patient and less powerful than they actually are. This seems to suggest the importance of making a

DOI: 10.2202/1559-0410.1296

protection adjustment as well. Finally, the main implication of this study is simply that protection does appear to exist, and suggestions by media, players, and managers that protection plays an important role in the game should not be dismissed so quickly.

References

[1] Angrist, J. and J.S. Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton.

[2] Bradbury, J.C. and D. Drinen (2008) 'Pigou at the Plate: Externalities in Major League Baseball.' *Journal of Sports Economics*, 9.

[3] Duffey, G. (2010) 'Riggleman Shuffles Lineup Order' www.nationals.com, 06/03/10.

[4] Duflo, E., et. al. (2004) 'How Much Should We Trust Difference in Differences Estimates?' *Quarterly Journal of Economics*, 119(1).

[5] Grabiner, D. (1991) Mimeo, http://www-math.bgsu.edu/ grabine/protstudy.txt

[6] James, B. (1985) *The Bill James Baseball Abstract, 1985* Ballantine Books.

[7] James, B. (2005) 'Underestimating the Fog.' *Baseball Research Journal*, 33.

[8] Lahman, S. (2009) 'Sean Lahman's Baseball Archive: Data from 1871-2009,' URL http://www.baseball1.com/, online resource.

[9] Moulton, B. (1990) 'An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables in Micro Units.' *Review of Economics and Statistics*, 72.

[10] Muskat, C. (2010) 'It's In With the New But Same 'Ol From Cubs' www.cubs.com, 05/05/10.

[11] Tango, T., et al (2007) *The Book: Playing the Percentages in Baseball* Potomac Books, Dulles.