MLB Team — Automatic Strike Zone

Final Analysis and Conclusions

Devon Goetz, Zack Kopstein, Simran Pabla, Claire Wichman

Introduction

In Major League Baseball, any play that transpires is ultimately at the discretion of the umpire. Many calls are obvious and clear, and the umpire plays little to no role in the outcome of the event. But inevitably, this is not always the case. What happens when the umpires make mistakes?

The MLB relies on humans to officiate all of the games, but humans are far from perfect. They carry inherent biases, are subject to external influences, and sometimes just make simple mistakes. We analyzed the difference these mistakes, when made at the plate, can make on the outcome of an inning. As we began to understand the impact these imperfections are having, we created a corrective and predictive algorithm to see how a game might transpire were it to be called perfectly. We then compared our manufactured "correct" games to the ones that had actually occurred, allowing us to draw conclusions about what a perfectly called game might look like, and how it might differ from reality.

The MLB currently possesses the ability to implement an Automatic Strike Zone, eliminating the need for an umpire to be calling balls and strikes. Before using it, however, they want to know how this addition might change the game, or how it could be implemented to most closely mimic the good habits of human umpires, while eliminating the costly mistakes they might be making.

Methodology

We approached this problem by breaking it down into smaller steps. We began by simply classifying mistakes. In order to do this, we separated pitches into four categories: correctly called balls, correctly called strikes, balls that were called strikes, and strikes that were called balls. Balls that were called strikes we labeled as defensive mistakes, because it is advantageous for the pitcher to have an extra strike, while we labeled strikes that were called balls as offensive mistakes.

Next, we created a new view in the Google Cloud dataframe for the mistake categorization, and from there were able to sum the number of mistakes made in an inning. Our first analytical approach was to examine the outcome of an inning, and see how those outcomes may vary depending on the number of mistakes made. Our four primary metrics are:

(1) Number of walks
(2) Number of strikeouts

(3) Number of runs scored
(4) Number of pitches thrown (as a way to estimate game time and duration)

We summed these for each inning, and put them into an additional column in our new dataframe view as well. Once we had this information, we were able to compare the average outcomes of innings with no mistakes, and compare them to a gradient of incorrect innings, trying to determine the influence of how many mistakes were made. We also broke it down one step further, and compared flawed at bats to at bats that had been called without mistakes. This allowed us to see a more immediate impact of the mistakes. We wanted to explore the occurrence and influence of mistakes more, so we looked at metric comparisons in specific situations, like when there were two outs, or in extra innings.

Knowing the effect umpire calls had on games that had already transpired was valuable in knowing the difference these mistakes were making. The next step we took was to create a model that allowed us to walk through a game and actually correct the mistakes that were made. In order to do this, we created a transition matrix that tabulated the probability of moving from any given state, as defined by the count, number of outs, run differential, and runners on base, into the next state. When our algorithm identifies a mistake, it creates a new, corrected state. For example, if the count is 2-0 and a pitch was mistakenly called a ball, the count moved to 3-0. We identify the mistake, and correct the count to 2-1. Using this new state, we calculate the most likely next state to occur, and proceed through. This becomes a more difficult prediction when the at bat ends with a different outcome than what had really happened. For example, if there was a ball mistakenly called a third strike, meaning that mistake caused the batter to strikeout, when we correct the count and predict a new outcome, that batter may not get out. If this is the case, we have another runner on base and one less out in the inning. Then we iterate through more predictive states until we reach the same number of outs as the next real batter should have when they stepped up to the plate. When we've reached the desired out state, we use the results from the remaining real plate appearances in the inning, with new runner positionings. Based on the real outcomes of these at bats, we probabilistically calculate the movement of the runners and adjust them accordingly.

With these new "corrected" games, we tallied the same metrics as before (strikeouts, walks, runs, and pitches) in order to compare the games that actually transpired and the corrected games that we manufactured. This gives us insight into the way an automatic strike zone could affect the games and how baseball would look were there to be no mistakes behind the plate.

Results and Discussion: Existing Inning Comparisons

To begin, Figure 1 displays how many mistakes are actually happening. These columns represent the number of innings that occured with the number of mistakes indicated on the x axis. The figure tallies the total number of mistakes of an inning; for example, there were 67,174 innings with no mistakes, and 792 innings with four mistakes. It also splits these mistakes into offensive and defensive mistakes; there were 19.120 innings that had one offensive mistake, and 38,012 innings that had one defensive mistake.

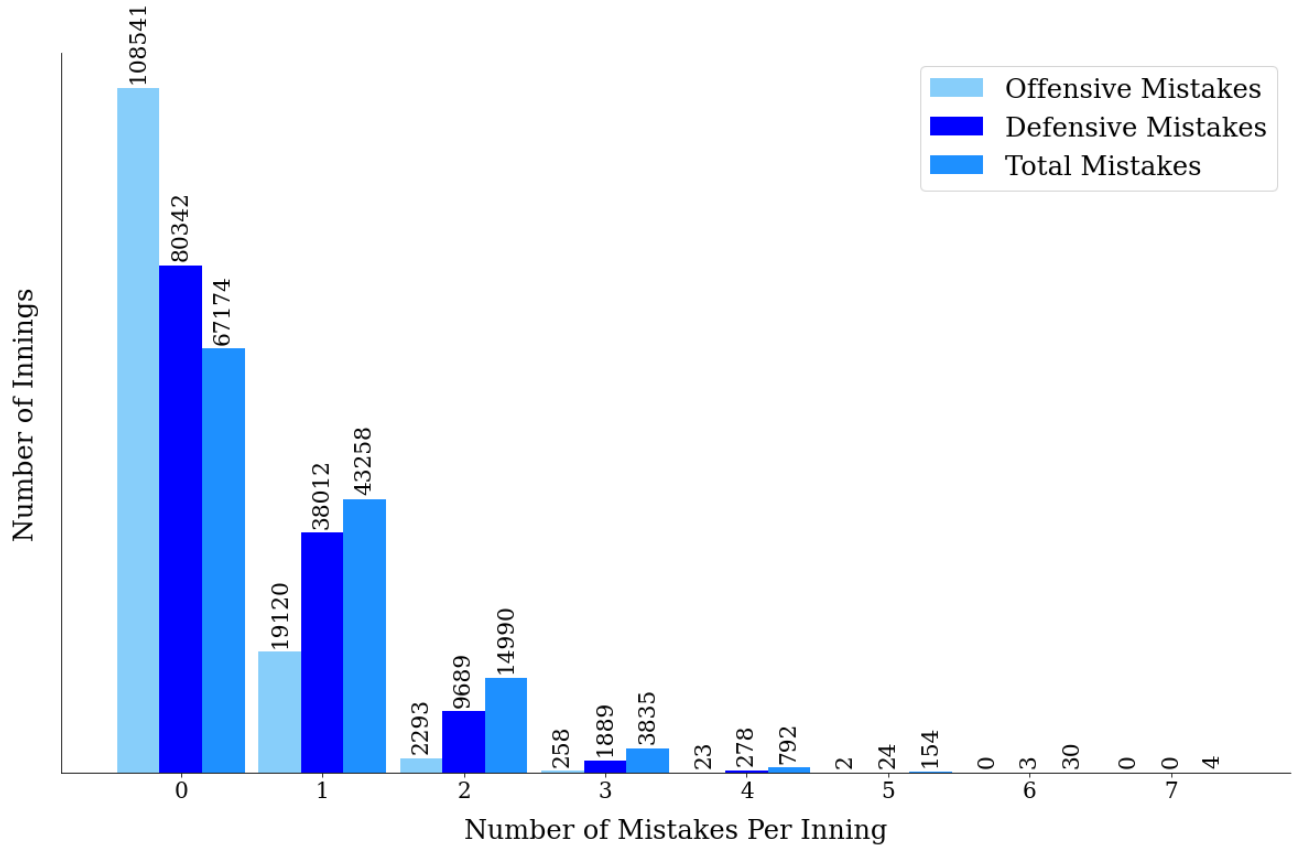


Figure 1. A bar chart that displays the total number of innings that have a certain amount of mistakes. There are three categories: offensive, defensive, and total. There is a clear downward trend, and the majority of the innings have between zero and two mistakes.

Offensive mistakes are the least common, as there are the most innings with zero offensive mistakes, and the offensive mistake bar is well below the others in all other categories. Overall, Figure 1 would indicate that umpires generally do a good job at correctly calling balls and strikes. When they are messing up, it is more often in the favor of the defense, meaning pitches outside of the strike zone are being called strikes.

In order to determine if certain parts of the game were subject to an increase in the number of erroneous calls, the number of mistakes in each inning number were tallied and shown in Figure 2. In order to account for fewer ninth innings, as the bottom of the ninth is not always played, and fewer extra innings, the number of mistakes was averaged by the number of times that inning number occured. For instance, there were 9,600 total mistakes made across all 5th innings (top and bottom), and there were 14,535 instances of 5th innings, meaning there were an average of 0.66 mistakes in each 5th inning. There were only 12 and 4 instances of the 18th and 19th innings, respectively, so the averages for those two, and the other higher inning numbers are less balanced.

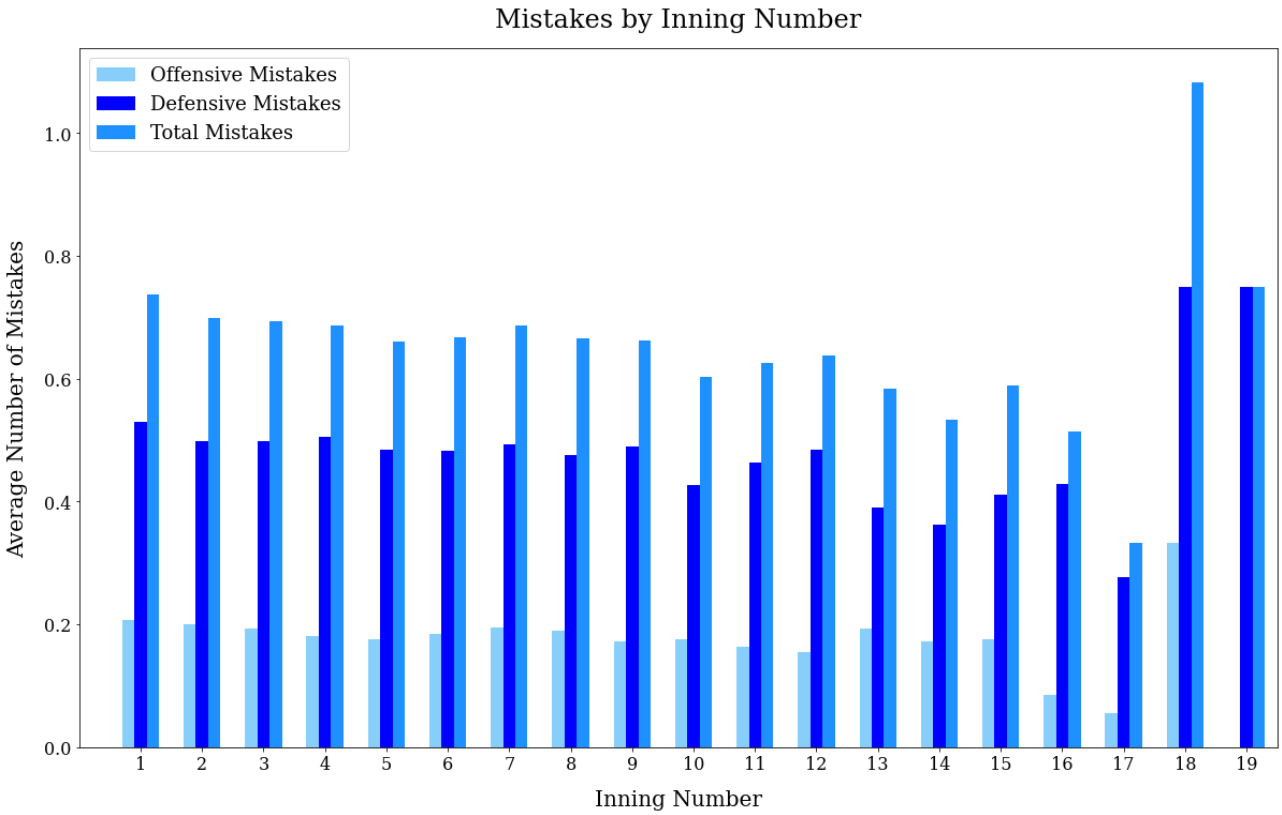


Figure 2. A bar chart showing the average number of mistakes made in each inning number. The total number of mistakes made in any given inning was divided by the instances of that inning. The final innings have fewer instances, so their averages are more variable. Overall. The number of mistakes per inning remains fairly consistent, with a slight downward trend as time progresses.

Across the first nine innings, the amount of mistakes stays relatively even, as do the number of offensive and defensive mistakes. There exists an extremely slight downward trend, indicating that the umpires make fewer mistakes as the games go on. We've hypothesized that this could be because they're getting more settled in and comfortable with the game, and also because the stakes are typically getting higher as the end of the game approaches.

To analyze the actual impact these mistakes were having on the game, we ran a comparison between innings with mistakes and innings with no mistakes, to see where the metrics differed. The first comparison we did was with **net mistakes**, or the number of offensive mistakes minus the number of defensive mistakes. In theory, if the two are weighted equally, the net categorization would be sorting innings by their general offensive or defensive advantage.

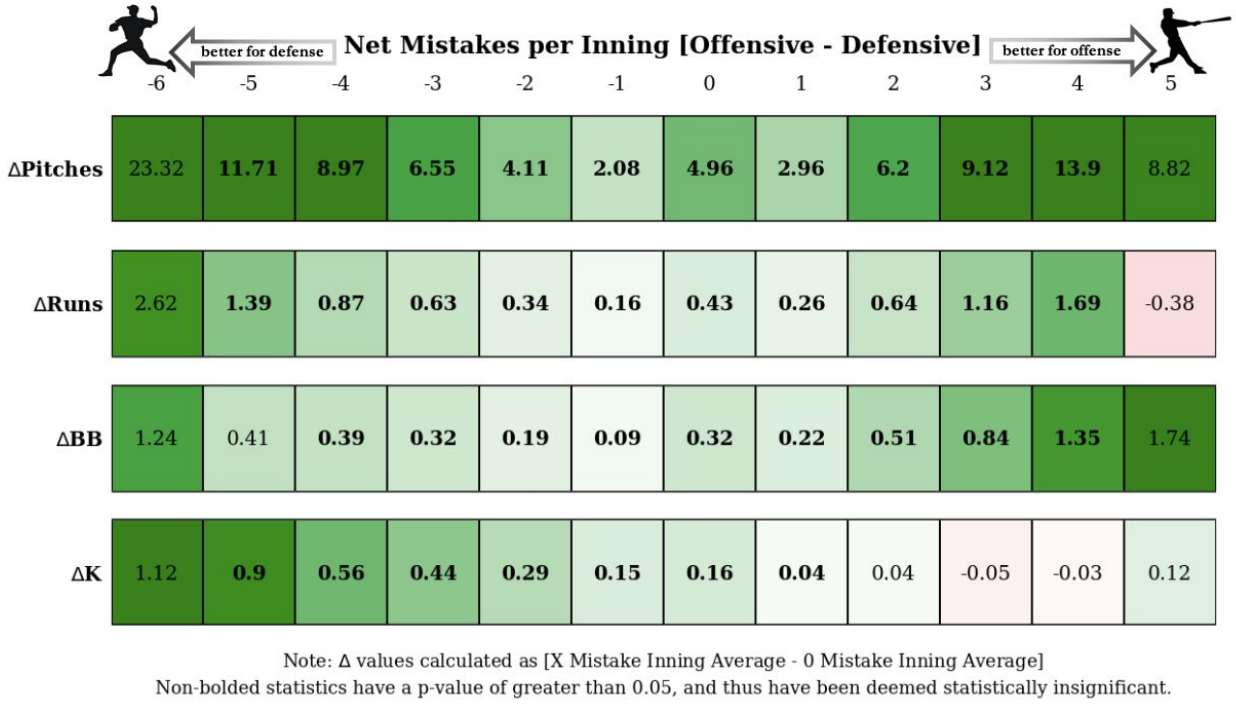


Figure 3. Along the x-axis, we have the net mistakes per inning, calculated as the number of offensive mistakes minus the number of defensive mistakes. For each metric, the number in the cell is the average for that metric across innings with the corresponding number of net mistakes, subtracted from the average for that metric across innings with no mistakes. Note that the zero column does not include innings with no mistakes in its averages, only innings with a net zero count.

Each row takes a generally parabolic shape, as can be seen in Figure 4, which represents the information in the table graphically. This is a relatively unexpected result; we had hypothesized that as more defensively advantageous mistakes were made, we would see fewer pitches, runs, and walks, while seeing an increase in strikeouts. Instead, we saw most metrics increase as we depart from the center, regardless of which side had more mistakes in their favor. One exception is strikeouts, which does more readily decrease as the offensive advantage increases, but the last four data points are not statistically significant, so only the first half of that trend is viable.

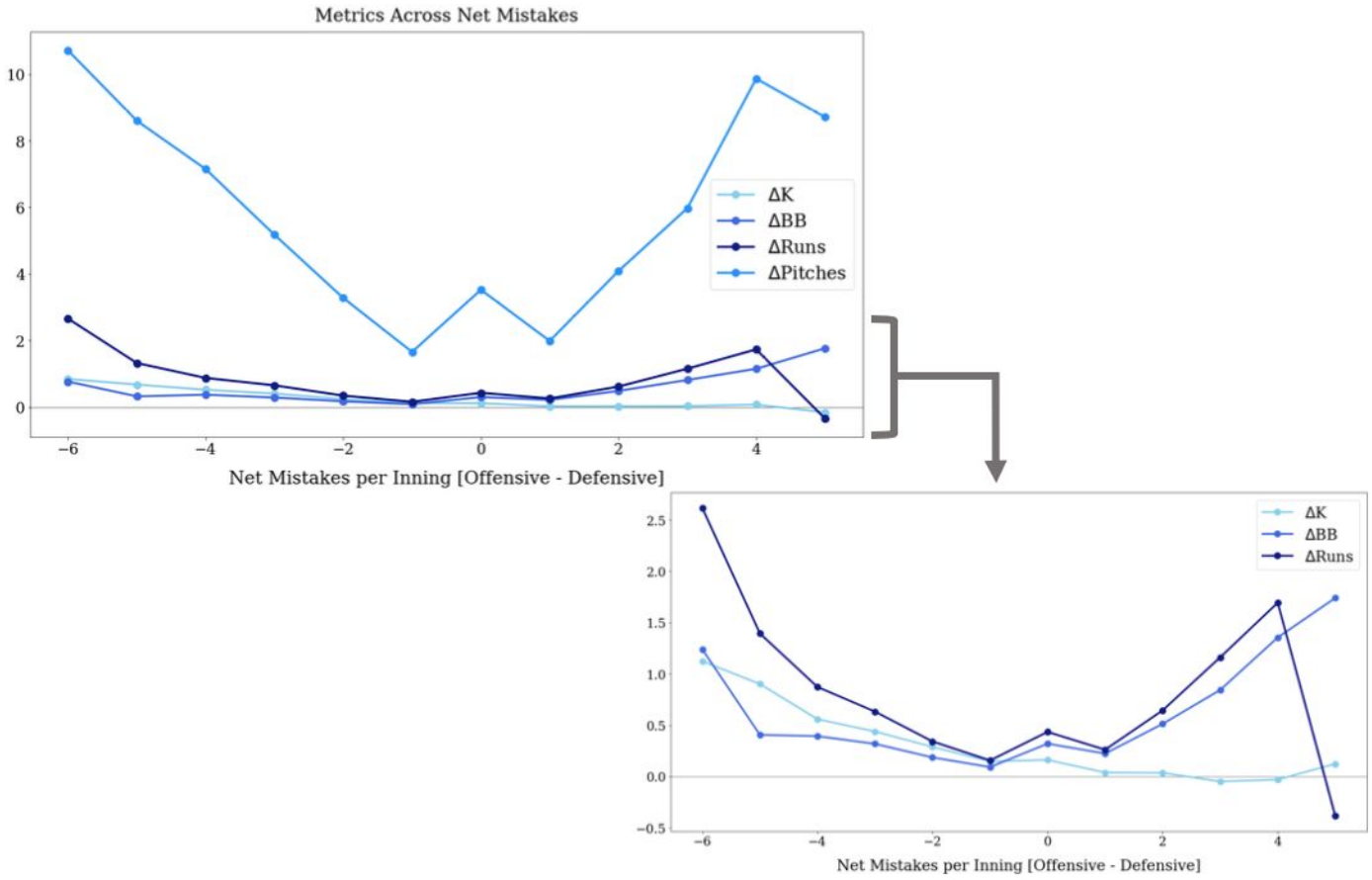


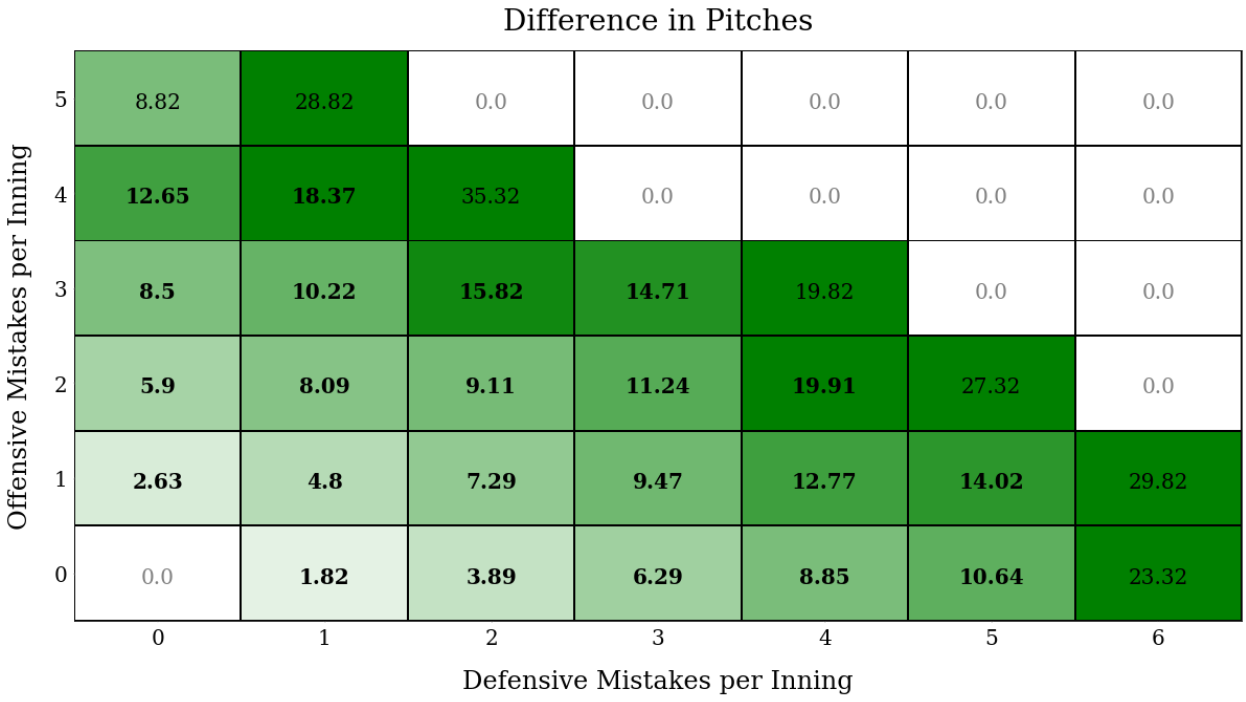
Figure 4. A graphical representation of the tabular data from Figure 3. It makes clear the parabolic trend we see in most of the metrics, a result that contradicted the original hypothesis of mistakes being advantageous. The notable exception is strikeouts, which does have a downward trend as the offensive mistakes increase. However, the last four data points, as noted by the asterisks in Figure 3, are not statistically significant.

We then delved into a deeper look, as it was unclear if these parabolic trends were due to the fact that we had grouped by net mistakes or some other underlying factor. Our main hypothesis was simply that the number of pitches was driving all metrics, including mistakes, instead of the other way around. Any pitch that has the opportunity to be called incorrectly is not a ball that is in play, and thus naturally drives the pitch count up, so the nature of the analysis skews that metric. In addition, the innings that last longer have more pitches thrown, and thus contain more opportunities for mistakes to be called, as well as the opportunity for more runs to be scored.

Figure 5 displays four tables, each corresponding to one of our metrics of focus, that groups the innings based on the combination of offensive and defensive mistakes in the inning. We see a similar trend as displayed in the Net Mistakes table, where regardless of the assumed offensive or defensive advantage, all metrics generally trend upward. The conclusion reached was that our

hypothesis was confirmed, that the number of pitches was driving all metrics, including mistakes, upwards.

A

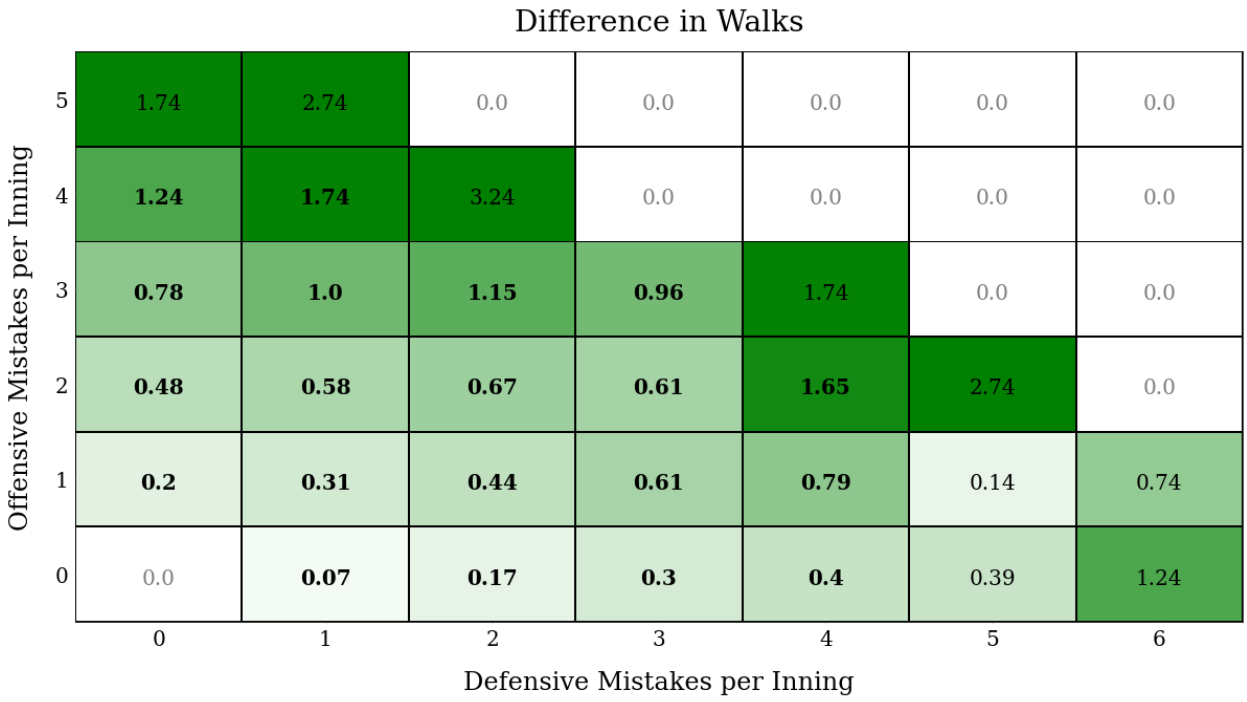


Note: Δ values calculated as [X Mistake Inning Average - 0 Mistake Inning Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

B

Note: Δ values calculated as [X Mistake Inning Average - 0 Mistake Inning Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

C



Note: Δ values calculated as [X Mistake Inning Average - 0 Mistake Inning Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

D

Note: Δ values calculated as [X Mistake Inning Average - 0 Mistake Inning Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

Figure 5. Each of these tables groups innings by the combination of offensive and defensive mistakes, and then compares the average metrics in those innings with the

averages in perfect, mistake-free innings. (a) This table shows the average number of pitches, again confirming that no matter the offensive or defensive categorization of mistakes, the number of pitches increases.

Because these four tables confirmed that the number of pitches was the driving factor, as well as that the analysis was biased by counting pitches from a study which fundamentally requires an uptick in pitch count each time, we decided to analyze on a smaller scale. Instead of grouping innings by their combinations of mistakes, we grouped by at bat in order to see the impact the mistakes were having on the outcome of an individual batter's opportunities at the plate. Figure 6 displays the net mistake table as well as the trendlines, and the specific combinations of offensive and defensive mistakes are tabulated in Figure 7.

A

Net Mistakes per At Bat [Offensive - Defensive]

	-3	-2	-1	0	1	2	3	4
ΔPitches	**1.22**	**1.4**	**0.91**	**1.75**	**1.14**	**1.62**	**1.83**	0.25
ΔRuns	**-0.11**	**-0.09**	**-0.04**	**-0.05**	**-0.02**	**-0.04**	0.03	-0.13
ΔBB	**-0.07**	**-0.02**	**0.01**	**0.12**	**0.14**	**0.37**	**0.77**	0.93
ΔK	**0.8**	**0.42**	**0.16**	**0.12**	**-0.01**	**-0.08**	**-0.2**	-0.2

Note: Δ values calculated as [X Mistake Inning Average - 0 Mistake Inning Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

B

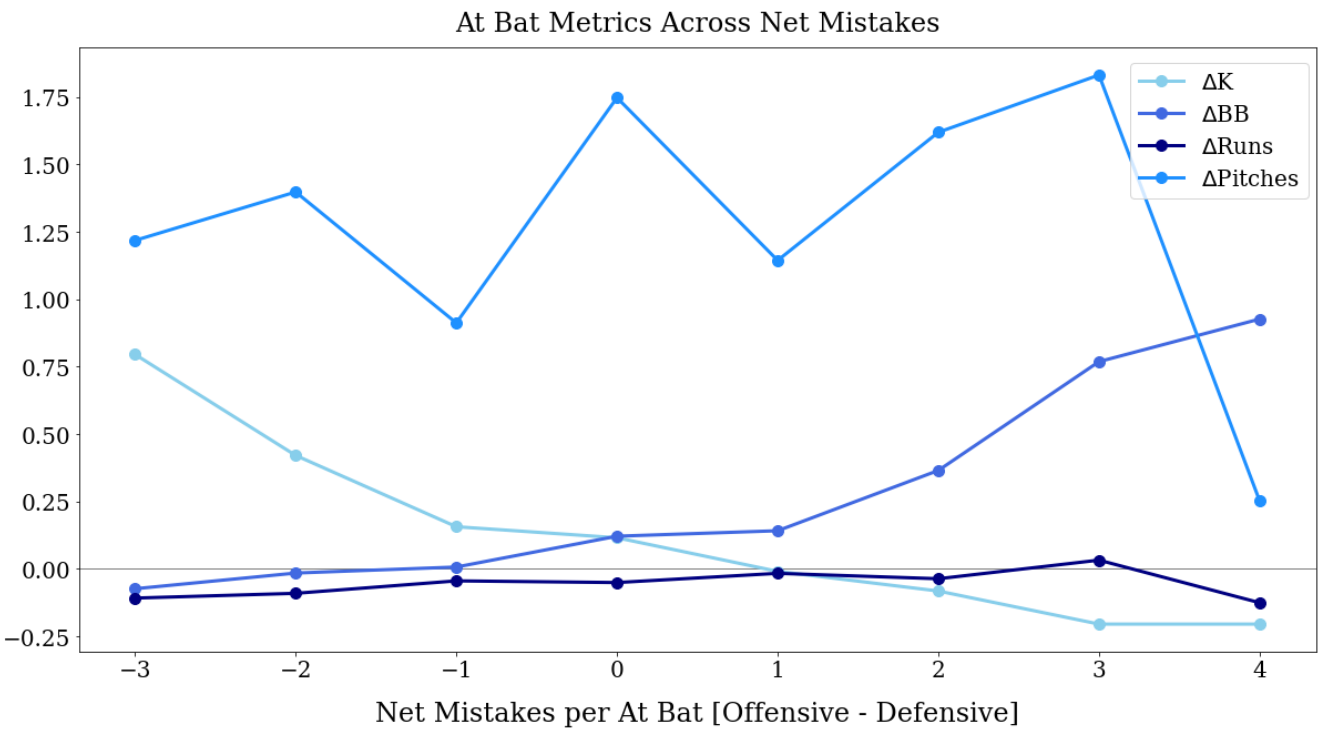


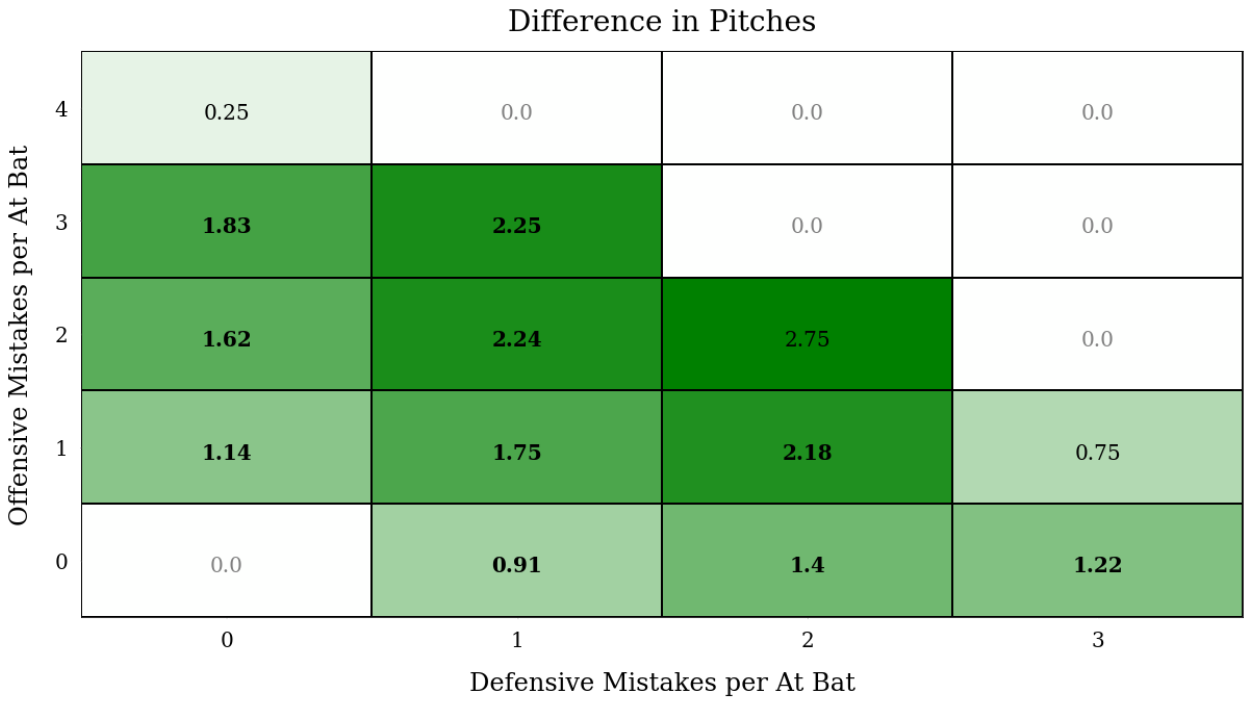
Figure 6. (a) This table shows the net number of mistakes and how that number impacts the metrics. As the net mistakes become increasingly positive, meaning more offensively advantageous, the number of strikeouts decreases, while the number of walks increase. Both pitches and runs follow weaker trends. **(b)** The line graph layout of the table allows us to see the trends more clearly, with strikeouts and walks trending in opposite directions. Runs increase very slightly as the offensive mistakes increase, while pitches demonstrate no real trend.

Across the net mistakes, the number of pitches appears to have very little trend, and the only real pattern is that every number is positive, even when perhaps it might be expected that the number of pitches would decrease if the defense was getting an advantage. This confirms the hypothesis that the analysis itself is biased against this metric, because any pitch logged as a mistake was not a pitch hit into play, forcing the pitch count upwards. Runs also display very little change, with an extremely slight upward trend as net mistakes increase and we lean towards an offensive advantage. The last data point is more dramatically negative, but is not statistically significant.

The two metrics of that display the clearest trends are walks and strikeouts. As net mistakes increase, walks mirror that trend, while strikeouts oppose it. As umpires make mistakes in favor of the offense, there is a clear uptick in the number of batters who walk, and as umpires make mistakes in favor of the defense, there is similar uptick in the number of batters who strikeout. Looking at Figure 6b, it can be concluded that the mistakes in one teams favor has a greater impact than the mistakes against a team. When there are more offensive mistakes, or net positive, there are a great deal more walks, but not many fewer strikeouts. There certainly are fewer strikeouts, but the

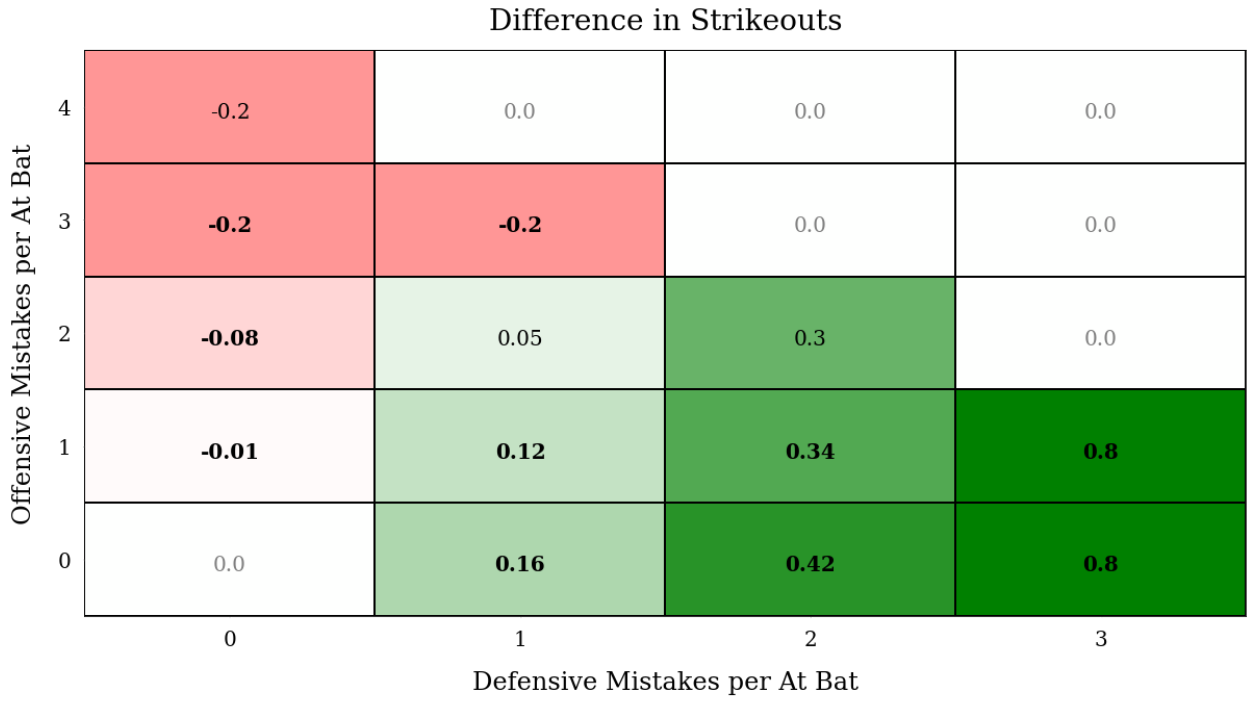
difference is of a much smaller magnitude than the difference in walks. Similarly, when there are more defensive mistakes, there is a larger increase in strikeouts than there is a decrease in walks. In order to look more closely at exactly how the mistakes are changing these metrics, the at bats were grouped by their unique combination of offensive and defensive mistakes.

A



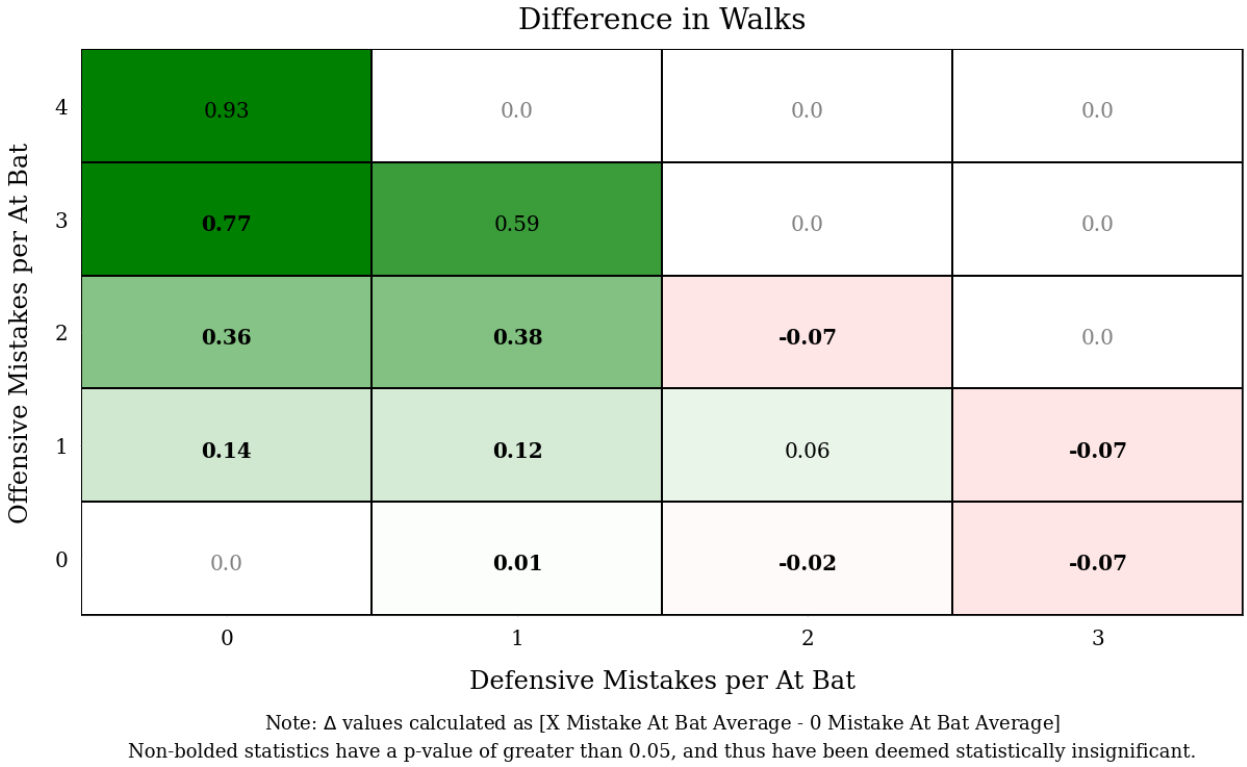
Note: Δ values calculated as [X Mistake At Bat Average - 0 Mistake At Bat Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

B



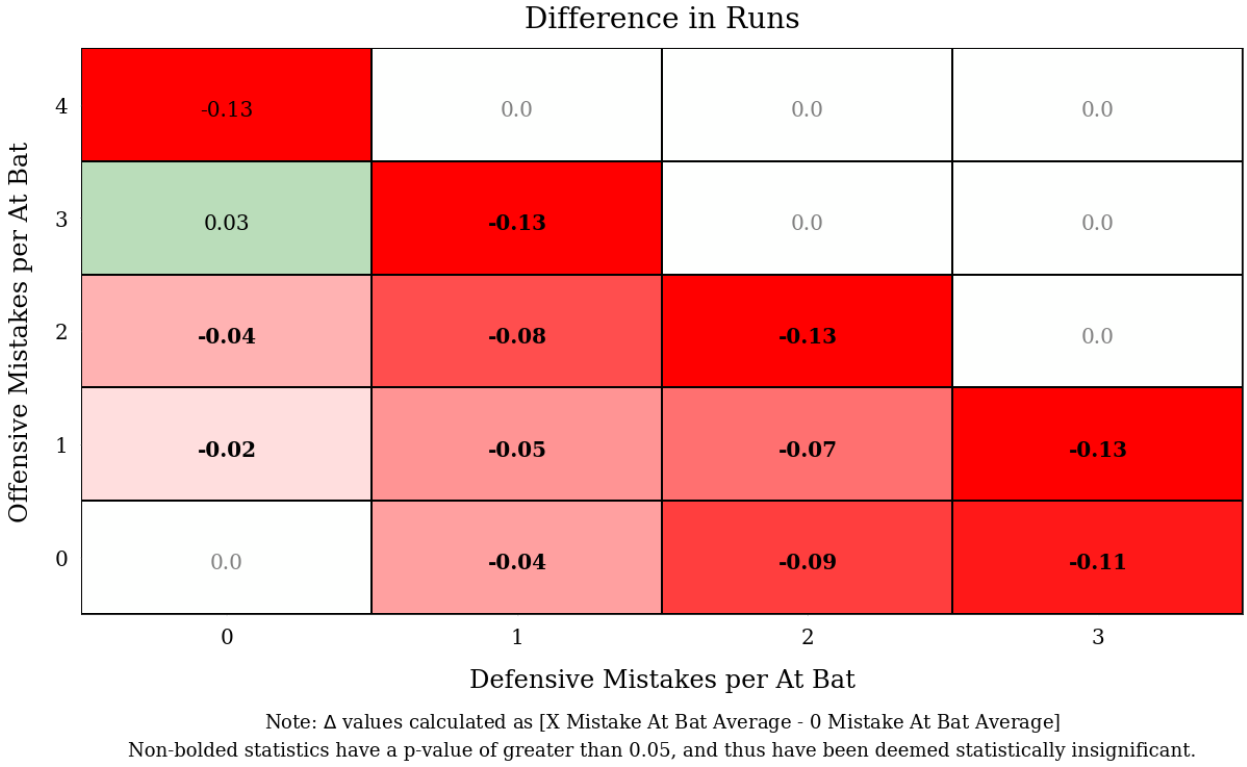
Note: Δ values calculated as [X Mistake At Bat Average - 0 Mistake At Bat Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

C



Note: Δ values calculated as [X Mistake At Bat Average - 0 Mistake At Bat Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

D



Note: Δ values calculated as [X Mistake At Bat Average - 0 Mistake At Bat Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

Figure 7. Four tables, each a breakout of one metric and how that fluctuates as the number of offensive and defensive mistakes change. **(a)** The number of pitches increases regardless

of mistakes, a byproduct of the fact that in order for a mistake to be tallied, a pitch that is not hit has to occur, automatically increasing the pitch count. **(b)** The number of strikeouts decreases as more offensively advantageous mistakes occur and increase as the number of defensively advantageous mistakes occur. **(c)** The opposite trend is seen in walks; they increase with more offensive mistakes and decrease with more defensive mistakes. **(d)** As perhaps the most puzzling trend, the number of runs scored decreases no matter the amount of mistakes that occur.

The difference in pitches was still an overall positive one, with values getting larger as the total number of mistakes increases, an expected outcome again due to the nature of the analysis.

At (1, 1) on the Strikeout Table, there is a positive delta value, perhaps suggesting that defensive mistakes play more strongly in influencing the outcome of a strikeout. As expected, the trends get stronger as we move towards more uneven at bats, with predominantly offensive or defensive mistakes.

Defensive mistakes play much less of a role in walks, bringing the average in a negative direction, but only by 0.07. By comparison

The number of runs scored is the most surprising trend. No matter the kind of mistake, runs seem to decrease when compared to perfect at bats. However, defensive mistakes do represent a larger decrease in runs, perhaps suggesting some slight impact on at bat outcome.

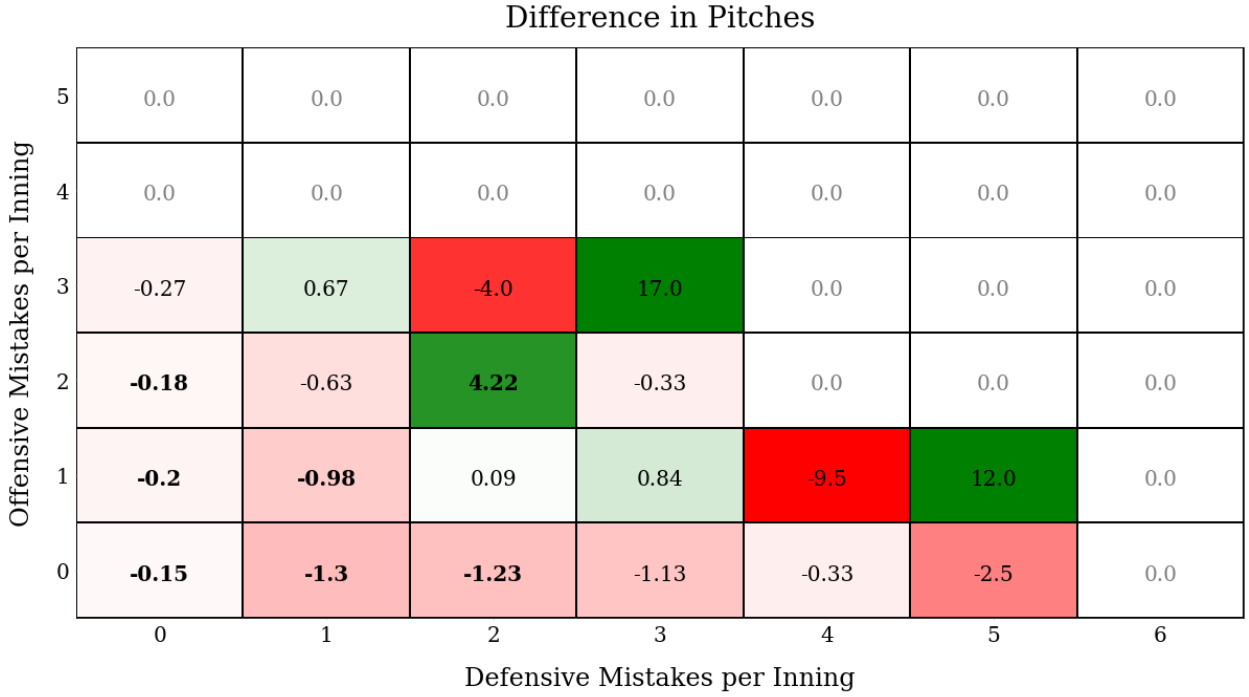
Results and Discussion: Corrected Inning Comparisons

Following the comparison between existing innings, we compared original innings to their "perfect" counterparts that our predict model produced.

The results for now are extremely preliminary, and are mainly here to placehold for real results. Our model is in its first iteration, so is not yet correcting offensive mistakes and we haven't incorporated some new data we just received that will help us be more accurate.

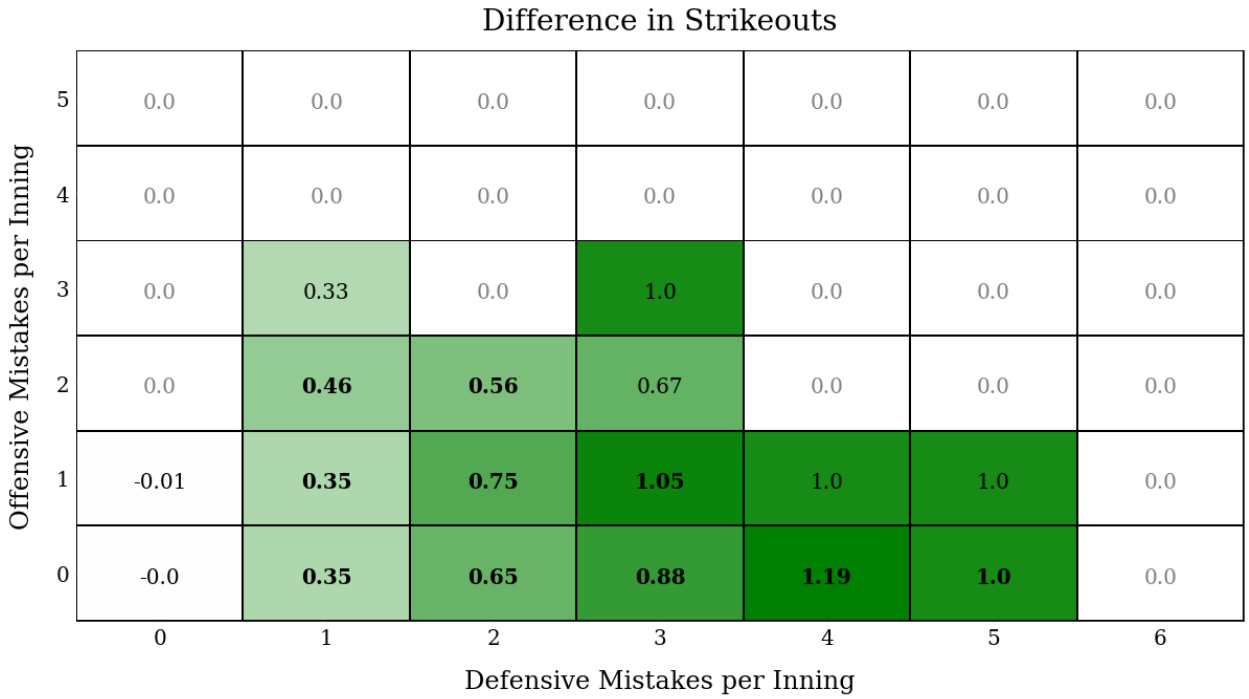
Figure 8 displays the difference we saw across our four targeted metrics.

A



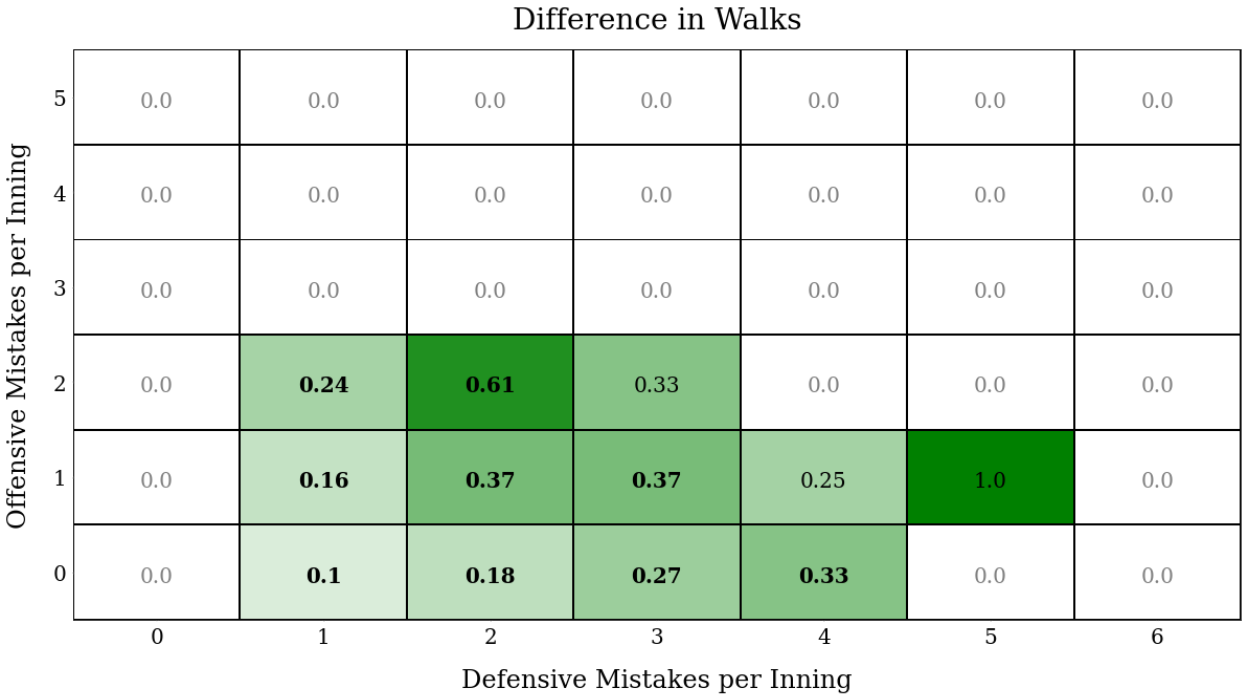
Note: Δ values calculated as [Corrected Inning Average - Original Inning Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

B



Note: Δ values calculated as [Original Inning Average - Corrected Inning Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

C



Note: Δ values calculated as [Original Inning Average - Corrected Inning Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

D

Note: Δ values calculated as [Original Inning Average - Corrected Inning Average]
Non-bolded statistics have a p-value of greater than 0.05, and thus have been deemed statistically insignificant.

Conclusions and Future Work

Throughout the course of this analysis, it has been determined that umpires are, in fact, making a real impact on the game. Their offensive and defensive mistakes help either team, and can impact a team's ability to produce or prevent offense.

One of the main things that remains to be done is to decide if this automatic strike zone is changing the game in a positive or negative way. This numerical analysis is just one piece of the puzzle, there are players, coaches, managers, fans, and many more people and factors to take into account before any final decision is made. Additionally, it may be that an automatic strike zone is the right choice, but how should it behave? Should it be the same size and shape? Should it stay consistent throughout the game? In order to determine that, more analysis of this kind would have to happen to see how those shape changes might influence the outcome of a game.

Acknowledgements

The authors would like to thank Peko Hosoi for her mentorship and support with every step of this project, as well as the rest of the 2.980 Staff for all of their help and support. They owe a huge thanks to Travis Buck and Reed MacPhail, their sponsors and mentors from the MLB for their helpful input and guidance, as well as the abundance of data they provided. The authors would also like to thank Devavrat Shah, Anish Agarwal, and Dennis Shen for their insights and guidance on causal inference and the predictive algorithm. They would like to thank Ramzi BenSaid for his assistance with their data sets and their Google interactions. Finally, they would like to thank the other students in the 2.980 class for their feedback and comments throughout the semester. None of this project would have been possible without all of these people and countless others. Thank you!