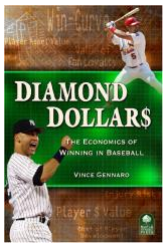

SABR—Diamond Dollars Case Competition

Measuring the Impact of Incorrect Ball—Strike Calls

Vince Gennaro
March 10, 2023
Case Competition

This case was prepared by Vince Gennaro and is developed solely for the purpose of a case discussion. It contains various assumptions that are generated for illustrative purposes and is not intended to serve as a source of primary data.

The data explosion in MLB has fundamentally changed baseball. It has revised our thinking on how runs are created, by influencing lineup construction and elevating the value of reaching base. The data proliferation may have done even more to support run prevention, by enabling pitchers to employ a data-driven approach to honing their craft. Pitch repertoires are optimized to the physical traits and the natural motion of the pitcher. Fielding has benefited from more precise positioning, turning more batted balls into outs.

Data has shined a bright light on another important area of the game, the home plate umpires ball-strike calls. Pitch tracking allows us to monitor the location of every pitch as it crosses (or not cross) the plate. Improvements in technology have led MLB to pilot an automated ball-strike system (ABS) on a broad scale in selected minor leagues in 2023, for potential adoption in 2024 or beyond. In the meantime, the baseball world, including umpires, will suffer through the transparency of "incorrect" calls, from the perspective of the ABS. The website umpscorecards.com regularly publishes the ball-strike accuracy of each home plate umpire for every MLB game, enumerating the correct-incorrect call of each pitch.

According to umpscorecards.com, the average number of incorrect calls in the 2438 regular and postseason MLB games in 2022 is 9.3. The most incorrect calls occurred on June 21, at Guaranteed Rate Field during a White Sox-Blue Jays game. Twenty-six of the 228 pitches were "called" incorrectly by umpire Doug Eddings, according to umpscorecards.com. Chicago won the game 7-6, begging the question, if all 26 missed calls were called correctly, how might it have affected the outcome of the game. The game with the least incorrect calls for 2022, occurred in the World Series. Umpire Pat Hoberg called a flawless game 2 of the Series, with 129 accurate calls and zero incorrect calls. Since beginning of the dataset in 2015, it is the only game of the 1900+ games recorded, where the umpires' ball-strike calls identically matched the automated system.

Your case problem is to analyze the way in which these missed calls might have affected the final score and outcome of selected games. For this case, I have identified 6 games from the 2022 season, all of which were decided by 1 or 2 runs. These 6 games contain various amounts of incorrect calls, ranging from 5 to 26. I've divided these games into two groups. Your assignment is to analyze at least one game from each group:

One from Group A & One from Group B

Group A

Score*	Date/Ump	Incorrect Calls
TOR 6 **CWS 7**	June 21 Eddings	26
NYY 6 **MIL 7**	Sept 16 Moscoso	22
OAK 7 **TOR 5**	April 16 Nelson	22

Group B

Score*	Date/Ump	Incorrect Calls
STL 2 **MIA 0**	April 20 Hallion	14
TEX 3 **MIA 2**	Sept 12 Libka	6
SFG 3 **ATL 4**	June 22 Hoberg	5

* Home team in **bold**

I have attached a png file from umpscorecards.com which shows the location of each incorrect call and provides a rudimentary estimate of the overall run impact of the missed calls, based simply on the change in run expectancy. I urge you to *not* rely on their estimate of the run impact, but rather to go beyond their approach to create your own methodology to assess the impact of the incorrect calls. To assist your analysis, I have attached csv files (from baseballsavant.com) that represent the detailed pitch logs for each of the 6 games. Your first task is to identify which pitches were called incorrectly. You will note that column AD (labeled "plate_x") represents the horizontal location of the pitch. A zero value in column AD means the pitch was down the center of the plate. (Negative values are towards the left-handed batters box and positive values are towards the right-handed batters box, from the pitcher's viewpoint) For the vertical strike zone, you can reference column AE (labeled "plate_z"), which represents the height of the pitch when it reached home plate. However, unlike the horizontal strike zone which is immovable, the vertical zone varies by batter (or even by pitch). So, you will need to calibrate the pitch against the top and bottom of the strike zone (columns AY and AZ, labeled "sz_top" and "sz_bot", respectively). Note: the measurements represent the center of the baseball, so do not ignore the width of the ball, when you identify which pitches were incorrect calls.

The Case Problem

Your goal is to effectively re-write history, by producing an alternative final score of two games—one from Group A and the other from Group B—based on every ball-strike call being correct. I am also asking you to provide some detail and commentary that highlights the changes for each individual game. This case problem presents a great deal of complexity. For example, if a 3-1 pitch was within the strike zone, but incorrectly called a ball, it leads to a walk. By correcting that call you will leave your batter with a 3-2 count, but with no further pitches in the plate appearance. This is an example where you will need to draw on your creativity. In effect, each incorrect call presents a "fork" in the flow of the game. How will you handle an incorrect called third strike that ends an inning, with multiple runners on base? Essentially, you are being asked to predict a game that never happened, but could have happened if each call was correct. You can take any approach to modelling the alternative outcome, as long as it is defensible.

Your output for this case assignment should be in the form of a powerpoint presentation to support a 20-minute oral presentation to a panel of judges, followed by a 10-minute Q & A by the judges. The judge's criteria will focus on the *quality of your analysis* and the *logic and creativity of your analytical framework* more than any single "right" answer. The ideal analysis has a logical flow, and is inclusive of the key factors that are expected to have resulted from eliminating all incorrect calls. More specifically, there are several key areas that will be a focus for the judges:

The presentation should include:

- **Your process**—a clear definition of the methodology used in developing your analysis and recommendations, including:
    - Is there solid logic that supports your approach?
    - The criteria you used to evaluate the data and draw conclusions
    - The statistical tools and techniques you employed
    - Your assessment of the risks associated with your approach

- **Your conclusions**—Does your analysis support your conclusions? You should also include the limitations of your analysis and acknowledge any risk factors.
- **Your creativity**—while you will not have time to go into detail on all of your analysis, did you think "outside the box" and address the problem you posed in a creative way and/or did you present your findings in a creative, effective way.
- **The quality and clarity of your presentation**—it's critical to carefully and strategically choose *what* to present and share with the judges. Storytelling is a critical aspect of influencing decisions through analytics.

A final comment regarding "rules" of the case and the competition:

- The intent of the competition is that team members are competing against other team members. This means that assistance from professors or non-members of the team is not permitted. Also, do not contact any MLB team or league personnel, or any other experts on non-experts, for advice on any of the case issues.
- You are encouraged to use the internet to help you with the case, particularly as a source of data, but be prepared to add your own insights, including quantitative analysis to the material you choose to draw from on the internet. One of the most common pitfalls for Case Competition participants is the over-reliance on analysis published on the leading analytical websites. While it is often valuable to consider these analyses, student teams have lost points by relying solely on these sites for answers to key case questions. We are looking to understand *your* analyses of the case questions, without an over-reliance on other peoples' thinking. For example, while umpscorecards.com provides a simplistic answer to the ultimate question we are posing in this case—the net impact of incorrect calls—you are expected to generate your own analysis and not simply rely on their answer.

I've provided an appendix below, which includes a glossary of terms of the column headings in the attached csv file

Appendix:

Statcast Search CSV Documentation

This is the documentation for the Statcast Search CSV data downloads.

pitch_type—The type of pitch derived from Statcast.

game_date—Date of the Game.

release_speed—Pitch velocities from 2008-16 are via Pitch F/X, and adjusted to roughly out-of-hand release point. All velocities from 2017 and beyond are Statcast, which are reported out-of-hand.

release_pos_x—Horizontal Release Position of the ball measured in feet from the catcher's perspective.

release_pos_z—Vertical Release Position of the ball measured in feet from the catcher's perspective.

player_name—Player's name tied to the event of the search.

batter—MLB Player Id tied to the play event.

pitcher—MLB Player Id tied to the play event.

events—Event of the resulting Plate Appearance.

description—Description of the resulting pitch.

spin_dir--* Deprecated field from the old tracking system.

spin_rate_deprecated--* Deprecated field from the old tracking system. Replaced by release spin

break_angle_deprecated--* Deprecated field from the old tracking system.

break_length_deprecated--* Deprecated field from the old tracking system.

zone—Zone location of the ball when it crosses the plate from the catcher's perspective.

des—Plate appearance description from game day.

game_type—Type of Game. E = Exhibition, S = Spring Training, R = Regular Season, F = Wild Card, D = Divisional Series, L = League Championship Series, W = World Series

stand—Side of the plate batter is standing.

p_throws—Hand pitcher throws with.

home_team—Abbreviation of home team.

away_team—Abbreviation of away team.

type—Short hand of pitch result. B = ball, S = strike, X = in play.

hit_location—Position of first fielder to touch the ball.

bb_type—Batted ball type, ground_ball, line_drive, fly_ball, popup.

balls—Pre-pitch number of balls in count.

strikes—Pre-pitch number of strikes in count.

game_year—Year game took place.

pfx_x—Horizontal movement in feet from the catcher's perspective.

pfx_z—Vertical movement in feet from the catcher's perpsective.

plate_x—Horizontal position of the ball when it crosses home plate from the catcher's perspective.

plate_z—Vertical position of the ball when it crosses home plate from the catcher's perspective.

on_3b—Pre-pitch MLB Player Id of Runner on 3B.

on_2b—Pre-pitch MLB Player Id of Runner on 2B.

on_1b—Pre-pitch MLB Player Id of Runner on 1B.

outs_when_up—Pre-pitch number of outs.

inning—Pre-pitch inning number.

inning_topbot—Pre-pitch top or bottom of inning.

hc_x—Hit coordinate X of batted ball.

hc_y—Hit coordinate Y of batted ball.

tfs_deprecated--* Deprecated field from old tracking system.

tfs_zulu_deprecated--* Deprecated field from old tracking system.

fielder_2—Pre-pitch MLB Player Id of Catcher.

umpire--* Deprecated field from old tracking system.

sv_id—Non-unique Id of play event per game.

vx0—The velocity of the pitch, in feet per second, in x-dimension, determined at y=50 feet.

vy0—The velocity of the pitch, in feet per second, in y-dimension, determined at y=50 feet.

vy0—The velocity of the pitch, in feet per second, in z-dimension, determined at y=50 feet.

ax—The acceleration of the pitch, in feet per second per second, in x-dimension, determined at y=50 feet.

ay—The acceleration of the pitch, in feet per second per second, in y-dimension, determined at y=50 feet.

az—The acceleration of the pitch, in feet per second per second, in z-dimension, determined at y=50 feet.

sz_top—Top of the batter's strike zone set by the operator when the ball is halfway to the plate.

sz_bot—Bottom of the batter's strike zone set by the operator when the ball is halfway to the plate.

hit_distance—Projected hit distance of the batted ball.

launch_speed—Exit velocity of the batted ball as tracked by Statcast. For the limited subset of batted balls not tracked directly, estimates are included based on the process described here.

launch_angle—Launch angle of the batted ball as tracked by Statcast. For the limited subset of batted balls not tracked directly, estimates are included based on the process described here.

effective_speed—Derived speed based on the the extension of the pitcher's release.

release_spin—Spin rate of pitch tracked by Statcast.

release_extension—Release extension of pitch in feet as tracked by Statcast.

game_pk—Unique Id for Game.

pitcher—MLB Player Id tied to the play event.

fielder_2—MLB Player Id for catcher.

fielder_3—MLB Player Id for 1B.

fielder_4—MLB Player Id for 2B.

fielder_5—MLB Player Id for 3B.

fielder_6—MLB Player Id for SS.

fielder_7—MLB Player Id for LF.

fielder_8—MLB Player Id for CF.

fielder_9—MLB Player Id for RF.

release_pos_y—Release position of pitch measured in feet from the catcher's perspective.

estimated_ba_using_speedangle—Estimated Batting Avg based on launch angle and exit velocity.

estimated_woba_using_speedangle—Estimated wOBA based on launch angle and exit velocity.

woba_value—wOBA value based on result of play.

woba_denom—wOBA denominator based on result of play.

babip_value—BABIP value based on result of play.

iso_value—ISO value based on result of play.

launch_speed_angle—Launch speed/angle zone based on launch angle and exit velocity.
1: Weak
2: Topped
3: Under
4: Flare/Burner
5: Solid Contact
6: Barrel

at_bat_number—Plate appearance number of the game.

pitch_number—Total pitch number of the plate appearance.

pitch_name—The name of the pitch derived from the Statcast Data.

home_score—Pre-pitch home score

away_score—Pre-pitch away score

bat_score—Pre-pitch bat team score

fld_score—Pre-pitch field team score

post_home_score—Post-pitch home score

post_away_score—Post-pitch away score

post_bat_score—Post-pitch bat team score

if_fielding_alignmentl—Infield fielding alignment at the time of the pitch.

of_fielding_alignment—Outfield fielding alignment at the time of the pitch.

spin_axis—The Spin Axis in the 2D X-Z plane in degrees from 0 to 360, such that 180 represents a pure backspin fastball and 0 degrees represents a pure topspin (12-6) curveball

delta_home_win_exp—The change in Win Expectancy before the Plate Appearance and after the Plate Appearance

delta_run_exp—The change in Run Expectancy before the Pitch and after the Pitch

This case was prepared by Vince Gennaro and is developed solely for the purpose of a case discussion. It contains various assumptions that are generated for illustrative purposes and is not intended to serve as a source of primary data.