PLAYER LATENT ABILITY THROUGH EVENTS (PLATE): SIMULTANEOUS PITCHER AND BATTER ABILITY MEASUREMENT WITH A NOVEL BAYESIAN ITEM RESPONSE THEORY TREE

George Perrett
New York University
gp77@nyu.edu

Klint Kanopka
New York University
klint.kanopka@nyu.edu

ABSTRACT

Current mainstream methods for evaluating pitcher performance in professional baseball do not yet account for differences in individual batter ability. Statistics like earned run average (ERA) and fielding independent pitching (FIP) are recorded for pitchers in both the major and minor leagues, however, low ERA or FIP in the minor leagues may not translate given the skill difference experienced when facing major league batters. Wins above replacement (WAR) is an alternative approach, but its primary utility is in identifying high ability players, not comparing players across the full ability spectrum. Additionally, WAR penalizes players who have played a small number of games and is not used to compare players within the minor leagues. We resolve these problems by introducing a novel methodology: Player Latent Ability Through Events (PLATE). PLATE is a Bayesian item response tree model that estimates pitchers latent ability as a function of different possible events at the end of a batters plate appearance (strike outs, field outs, runs batted in, hits, home runs etc). The primary innovation of PLATE is that the model formalizes player evaluation as a measurement problem and incorporates the difficulty of batters faced into estimates of a pitcher's underlying ability. This is done through simultaneous estimation of both pitcher and batter locations on a common scale, leading to not only estimates of pitcher ability, but interpretable parameters for individual batter characteristics and performance. The Bayesian estimation of PLATE provides both automatic uncertainty quantification and hierarchical regularization. Using publicly available data from all major league and Triple-A plate appearances in 2025, we show how PLATE facilitates managerial decision making by projecting major and minor league players onto a common scale, identifying elite pitching talent in fewer games than WAR, and highlighting meaningful ability variation among pitchers who appear similar under existing measures.

1 Introduction

Comparing and evaluating pitchers is critical to making informed decisions in professional baseball. Deciding which pitchers to call up or send down, or whom to trade for or trade away, can shape the outcomes of a season. From a baseball analytics perspective, a natural starting point is comparing existing measures of pitching performance, such as earned run average (ERA), Walks and Hits per Innings Pitched (WHIP), and Fielding Independent Pitching (FIP). Another approach is comparing Wins Above Replacement (WAR), an interpretable single number summary that has become a popular tool for player evaluation in professional baseball [1], American football [2], and hockey [3]. In professional baseball, WAR estimates how many additional major league wins a player contributes relative to an average minor league player at the same position.

A limitation of all of these statistics is that they do not place all pitchers onto a common scale and fail to account for the ability of batters faced. These limitations are most pronounced when attempting to compare pitchers across the major and triple a leagues. For pitchers who spend time in both the major leagues and triple A during a single season, their ERA, WHIP, and FIP are recorded separately for each league because these differing contexts make these statistics carry different meanings. Impressive ERA, WHIP, and FIP values in Triple A often fail to translate to major league performance, where pitchers face major league batters. Additionally, WAR is weighted by major league innings

pitched, making it particularly limited for comparing players who split time between the major league and Triple A and fundamentally unable to compare players who have not spent time in the major league.

Even when comparing players within the major league, existing publicly available pitching statistics do not account for the substantial variance in strength of schedule. Consider FIP, a statistic specifically designed not to punish pitchers for playing on teams with weak defense. Major league pitchers in the same division as Aaron Judge, one of the top power hitters in professional baseball, will be punished by FIP compared to pitchers in a division with fewer high ability power hitters. A pitcher striking out Aaron Judge or Luis Arraez, elite batters who have the lowest strikeout rates in the major leagues, should be weighted higher than striking out batters that frequently spend time in the minor leagues, however, the current publicly available measures of pitching performance treat these events as exchangeable indicators of a pitcher's underlying ability.

We address the limitations of current publicly available statistics used to compare pitchers by introducing Player Latent Ability Through Events (PLATE), a novel Bayesian Item Response Tree approach to pitcher evaluation. A primary innovation of PLATE is that our method formalizes pitcher evaluation as a latent variable measurement problem and leverages item response theory to account for differences in the quality of batters faced. PLATE provides a single number summary of a pitcher's latent ability that translates seamlessly across the major leagues and Triple-A. PLATE can generalize across leagues by leveraging players who split time between the major and minor leagues, allowing evaluation on a common scale. An ancillary benefit of PLATE is that the model simultaneously provides meaningful and interpretable information about batters and pitchers.

In this work, we estimate PLATE using every plate appearance from major league and Triple A baseball during the 2025 season. All data are publicly available through the `baseballr` R package [4, 5]. In the subsequent sections, we provide a brief introduction to item response theory, which theoretically informs PLATE. Present a detailed description of how PLATE estimates pitcher abilities and what information PLATE provides about batters. We show that PLATE can identify elite pitching talent much faster than existing measures such as WAR, and that PLATE's uncertainty quantification provides meaningful information about players' reliability and consistency.

2 Item Response Theory

We argue that pitching ability is a latent trait, meaning it cannot be directly observed, only inferred from discrete observable performances. High levels of pitching ability manifest in measures such as a high strikeout rate (K/9; strikeouts per nine innings), but these observable events are a downstream product of that pitcher's true, unobserved, latent ability. Item Response Theory (IRT) [6] provides a framework for the measurement of latent traits, such as math ability, depression, and pitching ability, from observable information, such as responses to math problems, clinical depression screeners, and strikeouts. IRT models also have a distinct advantage in baseball analytics, as pitcher abilities are linked together on a common scale based on overlap in the individual batters they have faced, naturally accounting for the ability of batters faced.

Consider the performance of Tarik Skubal during the 2025 season, where he was near unanimously considered the best pitcher in the American League [1]. Skubal produced an impressive K/9 of 11.10, but Dylan Cease and Garrett Crochet both posted more impressive K/9 values of 11.52 and 11.17, respectively. Jonah Tong, a Triple A pitcher, produced an even more impressive K/9 of 14.2. Does this imply that Cease, Crochet, and Tong have a higher ability of striking out batters than Skubal? Not necessarily. Skubal certainly faced stronger batters than Tong. It is also possible that Skubal faced stronger batters than Cease or Crochet, but this is harder to quantify, as it is highly dependent on team schedules.

The two parameter logistic IRT model (2PL), shown in Equation 1, can allow us to estimate a pitcher's latent ability from observed strikeouts. We denote a pitcher's true latent ability as θ_i, where i indexes all pitchers in the major leagues and Triple A during the 2025 season. The 2PL model estimates the probability that plate appearance X_{ij} where batter j faces pitcher i results in a strikeout. $X_{ij} = 1$ denotes the plate appearance ended in a strikeout while $X_{ij} = 0$ denotes that plate appearance did not end in a strike out.

In addition to pitcher ability parameters θ_i, this 2PL model estimates two batter parameters: batter *strength* (β_j), and batter *discrimination* (α_j). In the context of strikeouts, batter strength is the extent to which batters are difficult to strikeout. Consider Luis Arraez who took the fewest strikeouts in the major league during the 2025 season. We would expect Arraez to have one of the highest values of batter strength, the majority of his plate appearances did not end in a strikeout. Note that Equation 1 depends on the *difference* between pitcher ability and batter strength $(\theta_i - \beta_j)$. Because of the specific functional form, we expect that when $\theta_i = \beta_j$ the probability of the plate appearance ending in a strikeout is essentially a coin flip: $P(X_{ij} = 1 \mid \theta_i) = 0.5$. When pitcher ability is higher than batter strength, strikeouts become more likely, $\theta_i > \beta_j \rightarrow P(X_{ij} = 1 \mid \theta_i) > 0.5$, and vice versa. Meanwhile, batter discrimination informs us

[1] Skubal received 26 of 30 first place votes in the AL Cy Yong voting

about how sensitive a specific batter's performance is to the difference between a pitcher ability and batter strength. When $\alpha_j = 0$, the probability of strikeout does not vary with pitcher ability at all and $P(X_{ij} = 1 \mid \theta_i) = 0.5$. When a $0 < \alpha_j < 1$, the role of pitcher ability is diminished and larger ability differences are required to produce changes to strikeout probability. Similarly, when $\alpha_j > 1$, the role of pitcher ability is magnified and smaller differences produce larger changes in strikeout probability. A high discrimination batter would have strong performances against low ability pitchers but would crumble when facing elite pitching talent while a low discrimination batter would have consistent performance regardless of who is on the mound. An important consequence of this framing is that pitcher ability and batter strength are measured on the same scale.[2]

$$\Pr(X_{ij} = 1 \mid \theta_i) = \frac{1}{1 + e^{-\alpha_j(\theta_i - \beta_j)}} \tag{1}$$

As such, the use of a 2PL model for pitching ability allows us to place the strikeout ability of Skubal, Cease, Crochet, and Tong on the same scale while simultaneously accounting for differences in quality of the batters they faced. In fact, even if two pitchers have never faced a common batter, their estimated abilities are linked and comparable on the same scale if a path of pitcher and batter matchups of any length can be drawn between them. This property of IRT models allows for comparison between the American and National Leagues due to Interleague play and comparison between Major and Minor leagues because of players who are called up or sent down over the course of their careers. However, an obvious limitation of this particular model is that strikeouts are not the only manifestation of a pitcher's true ability. To fully capture the skillset of a pitcher we need to consider a more full range of events that characterize the interactions between a pitcher and batter.

3 Player Latent Ability Through Events (PLATE)

Item Response Trees (IRTrees) are a flexible class of IRT models that estimate latent ability from a sequential series of conditional events[7, 8]. We developed Player Latent Ability Through Events (PLATE) to consider numerous manifestations of pitching ability that may appear over the course of a plate appearance. Unlike the example 2PL model presented in the previous section, PLATE is able to simultaneously incorporate strikeouts, out accumulation, run prevention, and power hitting by leveraging the flexibility of IRTrees. While IRTree models have traditionally been applied to complex measurement problems in educational testing, psychology, and survey research, to our knowledge, PLATE the first application of an IRTree framework to baseball analytics or sports analytics more broadly.

PLATE works by dividing each plate appearance into a set of mutually exclusive conditional events organized into a decision tree. Every node in the tree is consistently coded such that, conditional on tree depth, higher values always represent better outcomes for pitchers and lower values always represent better outcomes for batters. The consistent coding of each node is essential for a meaningful interpretation of both pitcher ability parameters θ_i and batter parameters β_j and α_j. The structure of the PLATE Item Response Tree is shown in Figure 1. The top of the tree where a pitcher either strikes out a batter or does not strike out a batter is assigned the same 2PL functional form shown in Equation 1, however, unlike the simple 2PL model described earlier, the PLATE IRTree continuous beyond strikeouts and encompasses a broader set of events that defines an encounter between a pitcher and batter. The PLATE estimates presented here are fit from the 2025 major league and Triple A seasons and informed by a total of 367,383 unique plate appearances.[3] We provide pitcher ability estimates for 1,559 unique pitchers, and batter strength and discrimination estimates for 1,124 unique batters.

The PLATE IRTree splits plate appearances into those ending with a strike out or not ending with a strike out. Conditional on the plate appearance not ending with a strikeout, the plate appearance can lead to an out or not lead to an out. When an out occurs, it is possible that the out led to a surrendered run (RBI) or that the out did not surrender any runs (no RBI). For plate appearances that do not end in an out, a pitcher either surrenders a power hit (double, triple, or home run) or does not and the batter advances to first base (through either a single, walk, or being hit with a pitch). Non-power hits can still surrender runs (RBI), or end with no surrendered runs (no RBI). For power hits, we consider surrendered runs at a more granular level, no runs (No RBI), one run (1 RBI), two runs (2 RBI), or three or four runs (3+ RBI). For plate appearances that did not result in a power hit, we collapsed RBIs into a binary categories because it is rare that more than 1 run is scored without a power hit[4]. Likewise, power hits scored 3 runs 5.6% of the time and scores 4 runs 1.06% of the time so these categories were collapsed.

[2]While similar in feel and functional form to an Elo system, note that pitchers who also hit, such as Shohei Ohtani, would have separately estimated θ_i, β_j, as the comparison is cross-classified across two disjoint skill sets.

[3]Note that the PLATE IRTree excludes all plate appearances that ended with an out due to a failed attempt at stealing a base

[4]Given that a power hit has not happened more than 1 run was scored only 0.65% of the time.

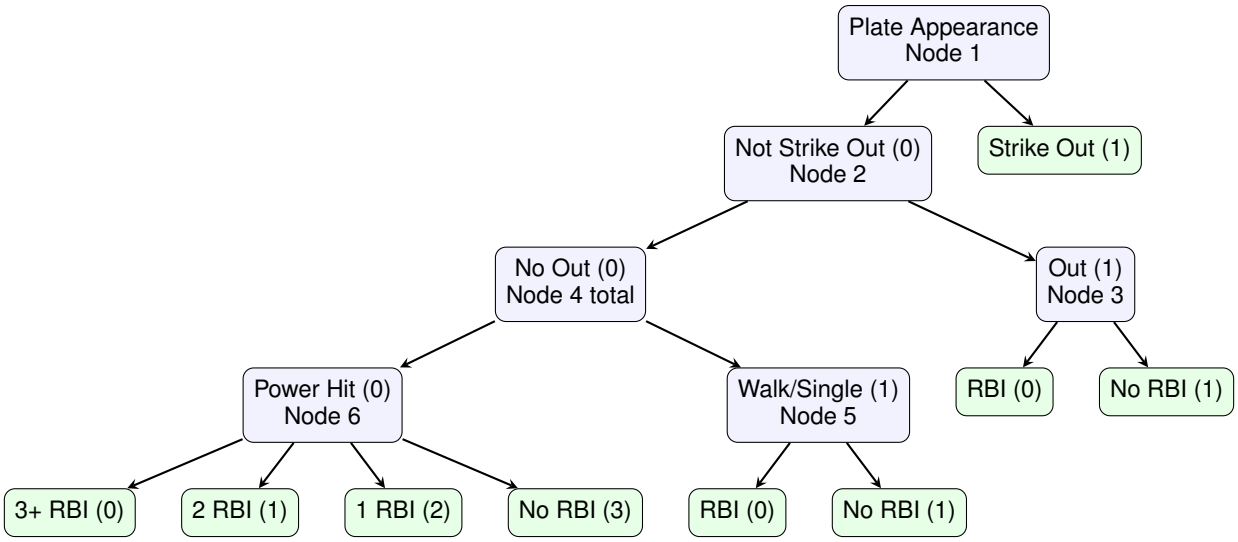

Figure 1: The PLATE Item Response Tree. Non-terminal nodes are shown in blue while terminal nodes are shown in green. Coded values of each node are presented within parentheses.

Each non-terminal node in the PLATE Item Response Tree is assigned an IRT model. Pitcher ability parameters θ_i are shared across IRT models while batter strength and batter discrimination (β_{jk} and α_{jk}, respectively) are unique to each tree node, with k indexing the node. This parameterization achieves the primary goal of developing PLATE: a single one number summary of pitching ability. While the batter parameters estimated by PLATE are interpretable and offer value, these parameters (strength and discrimination) are always contextually interpreted at the node level.

Equation 4 shows the likelihoods for the six non-terminal nodes in the PLATE IRTree. We use $X_{ij,k}$ to represent the observed event of node k. When $k = 1$, $X_{ij,k} = 1$ indicates that a plate appearance ended with a strikeout. For node 2, where $k = 2$, $X_{ij,k} = 1$ indicates that a plate appearance ended with an out, and so forth. Nodes $k \in \{1, \ldots, 5\}$, with binary outcomes are assigned 2PL IRT models. The structure of these 2PL models resembles the simple IRT model presented earlier in Equation 1. However, unlike the earlier 2PL model, we now also adjust for pitch count (see the expression for $\eta_{ij,k}$ in Equation 4) and measure pitchers' latent ability across the wider range of events encoded in the IRTree, rather than limiting our measure of pitcher latent ability to only strikeouts.

Node 6, RBIs conditional on a power hit, splits into 4 ordered categories. To account for the inherent ordering of outcomes in this node, we assign an ordered logistic likelihood, shown in Equation 4. From an IRT perspective, this is known as a Graded Response Model (GRM)[9]. For Node 6, there are four outcomes, $X_{ij,6} \in \{0, 1, 2, 3\}$. The GRM specifies the following conditional probabilities:

$$
\begin{aligned}
P(X_{ij,6} \geq 1 \mid \theta_i) &= \frac{1}{1 + e^{-\alpha_{j,6}(\theta_i - \tau_{j,1})}} \\
P(X_{ij,6} \geq 2 \mid \theta_i) &= \frac{1}{1 + e^{-\alpha_{j,6}(\theta_i - \tau_{j,2})}} \\
P(X_{ij,6} \geq 3 \mid \theta_i) &= \frac{1}{1 + e^{-\alpha_{j,6}(\theta_i - \tau_{j,3})}}
\end{aligned}
\tag{2}
$$

In our context, the GRM estimates thresholds, denoted as $\tau_{j,m}$, at which the result of a plate appearance has a probability of 0.5 of falling in category m or above. Conceptually, this is similar to fitting a series of 2PL models that divide the ordered result space in two at a threshold category. With 4 categories, the GRM estimates 3 thresholds. From this, it is also straightforward to compute the conditional probability of each individual result:

$$
\begin{aligned}
P(X_{ij,6} = 0 \mid \theta_i) &= 1 - P(X_{ij,6} \geq 1 \mid \theta_i) \\
P(X_{ij,6} = 1 \mid \theta_i) &= P(X_{ij,6} \geq 1 \mid \theta_i) - P(X_{ij,6} \geq 2 \mid \theta_i) \\
P(X_{ij,6} = 2 \mid \theta_i) &= P(X_{ij,6} \geq 2 \mid \theta_i) - P(X_{ij,6} \geq 3 \mid \theta_i) \\
P(X_{ij,6} = 3 \mid \theta_i) &= P(X_{ij,6} \geq 3 \mid \theta_i)
\end{aligned}
\tag{3}
$$

Besides pitcher ability θ_i and node specific batter parameters $\beta_{j,k}$ and $\alpha_{j,k}$, every IRT model within the PLATE Item Response Tree includes an additional pitch count parameter ζ_k. As with batter parameters, the k subscript denotes that there are six different pitch count parameters, one for each non-terminal node. This allows pitcher fatigue to differentially influence plate appearance outcomes. A starting pitcher giving up a home run in the 9th inning is less indicative of their latent ability than a starting pitcher who gives up a home run on the first pitch of the game. As such, we include the pitch count parameter to avoid punishing starting pitchers for going deeper into games. Pitch count is measured at the *end* of each plate appearance. As such, our full PLATE model estimates:

$$
\begin{aligned}
\eta_{ij,k} &= \zeta_k \cdot \text{pitchcount}_{ij} + \alpha_{j,k}\big(\theta_i - \beta_{j,k}\big) \\
\mathcal{L}_k &= \begin{cases} \text{Bernoulli}\big(X_{ij,k} \mid \text{logit}^{-1}(\eta_{ij,k})\big), & k = 1, \ldots, 5 \\ \text{OrderedLogistic}\big(X_{ij,k} \mid \eta_{ij,k}, \tau\big), & k = 6 \end{cases}
\end{aligned}
\tag{4}
$$

Estimating IRT models requires constraining parameter values. Without a constraint, IRT models are inherently unidentified, as adding a constant to each θ_i, $\beta_{j,k}$ and $\tau_{j,m}$ gives an equivalent scale. A separate estimation challenge unique to PLATE is that some batters have only a single plate appearance in major league or triple-A baseball during the 2025 season. For batters with few plate appearances, batter parameters $\beta_{j,k}$ and $\alpha_{j,k}$ will become unstable. We address both identification issues and small observation challenges by assigning prior distributions to the parameters and estimating PLATE within a Bayesian inference framework. Prior distributions are shown in Equation 5. Bayesian priors on IRT parameters provide soft constraints that typically resolve identification problems by regularizing estimates towards the mean of the distribution. Given the complexity of the PLATE and volume of available data, we also constrain the pitcher abilities to be mean zero ($\bar{\theta}_i = 0$), a hard identification constraint employed in frequentist IRT estimation. PLATE also induces a multilevel hierarchical structure on all batter strength parameters $\beta_{j,k}$ and batter discrimination parameters $\alpha_{j,k}$. Hierarchical priors induce regularization from partial pooling, such that batters with few plate appearances borrow strength from batters better represented in the data and are regularized towards average values of batter strength and batter discrimination.

$$
\begin{aligned}
&\theta_i \sim \mathcal{N}(0, 1),\ \textstyle\sum_i \theta_i = 0 \\
&\alpha_{j,k} \sim \text{LogNormal}(\mu_{\alpha,k}, \sigma_{\alpha,k}) \\
&\beta_{j,k} \sim \mathcal{N}(\mu_{\beta,k}, \sigma_{\beta,k}) \\
&\mu_{\alpha,k} \sim \mathcal{N}(0, 0.2),\ \mu_{\beta,k} \sim \mathcal{N}(0, 0.5)\ \ (k = 1, \ldots, 5) \\
&\sigma_{\alpha,k} \sim \text{Exponential}(1),\ \sigma_{\beta,k} \sim \text{Exponential}(1) \\
&\zeta_k \sim \mathcal{N}(0, 1) \\
&\tau_j \sim \mathcal{N}(0, 1),\ \tau_j \in \mathbb{R}^{K_{\text{rbi}}-1}\ \ \text{(ordered)}
\end{aligned}
\tag{5}
$$

PLATE estimates are the conditional probability of the parameters given the observed events across plate appearances. These conditional probabilities are estimated by applying Bayes theorem, giving us the posterior distribution shown in Equation 6. To fit the PLATE model we use the `rstan` package [10] to interface with Stan version 2.39.0 [11]. **Running the PLATE model requires Stan version 2.36 or later**.

$$
\begin{aligned}
&\phi_k = \{\mu_{\beta,k},\ \sigma_{\beta,k}\} \\
&\psi_k = \{\mu_{\alpha,k},\ \sigma_{\alpha,k}\} \\
&p(\theta_i, \beta_{j,k}, \alpha_{j,k}, \zeta_k, \tau_j, \phi_k, \psi_k \mid X_{ij,k}) \propto p(\theta_i)\, p(\tau_j) \prod_{k=1}^{6} \Big[\mathcal{L}_k\ p(\beta_{j,k} \mid \phi_k)\ p(\alpha_{j,k} \mid \psi_k)\ p(\zeta_k)\ p(\phi_k)\ p(\psi_k) \Big]
\end{aligned}
\tag{6}
$$

In our analysis of the 2025 major league and Triple A seasons we sampled from the posterior distribution with four independent chains each drawing 2000 samples from the posterior after completing 1000 warmup iterations. All Rhat values are below 1.01 and the smallest effective sample size is greater than 400 signaling good convergence.

4 Evaluating Pitchers with PLATE

Due to consistent coding across the PLATE IRTree, higher PLATE estimates denote higher pitching ability. One benefit of projecting major and minor league players onto a common scale is that PLATE can leverage a pitcher's plate appearances across both leagues in their ability estimates. This allows PLATE to identify and quantify high ability pitching, even among players with a limited number of major league appearances. Player evaluations from WAR and PLATE, two competing one number summaries of pitcher's performance, are presented in Figure 2.

A critical limitation of WAR is that it penalizes pitchers who have pitched fewer innings in the major leagues. Consider the young rising stars Trey Yesavage, Chase Burns, and Jacob Misiorowski, who all spent a considerable portion of the 2025 season in Triple A. Burns and Misiorowski were both called up to the major leagues midway through the 2025 season, and Yesavage was not called up until the last month of the season. Despite the late call up, Yesavage played a pivotal role for his team starting both Game 1 and Game 5 of the 2025 World Series. Likewise, Burns and Misiorowski made critical contributions to their respective teams during the 2025 season and would both go on to become Cy Young contenders during the subsequent 2026 season. PLATE correctly identifies the elite ability of these three pitchers, while WAR is a lagging indicator that systematically undervalues their ability during the 2025 season.

Comparing established starting pitchers Michael Wacha and Zach Wheeler demonstrates that WAR's prioritization of innings pitched can also affect established major league pitchers. Despite a year where he posted an elite ERA of 2.71 and was recognized as a Cy Young finalist, Wheeler has a comparable WAR to Wacha, who posted an ERA of 3.86. Wheeler and Wacha have comparable WAR ratings because Wheeler missed time during the 2025 season due to injury. WAR prioritizes major league innings pitched, so established players who miss time will be penalized. PLATE prioritizes performance on the mound and indicates that, conditional on availability, Wheeler is clearly the better pitcher.

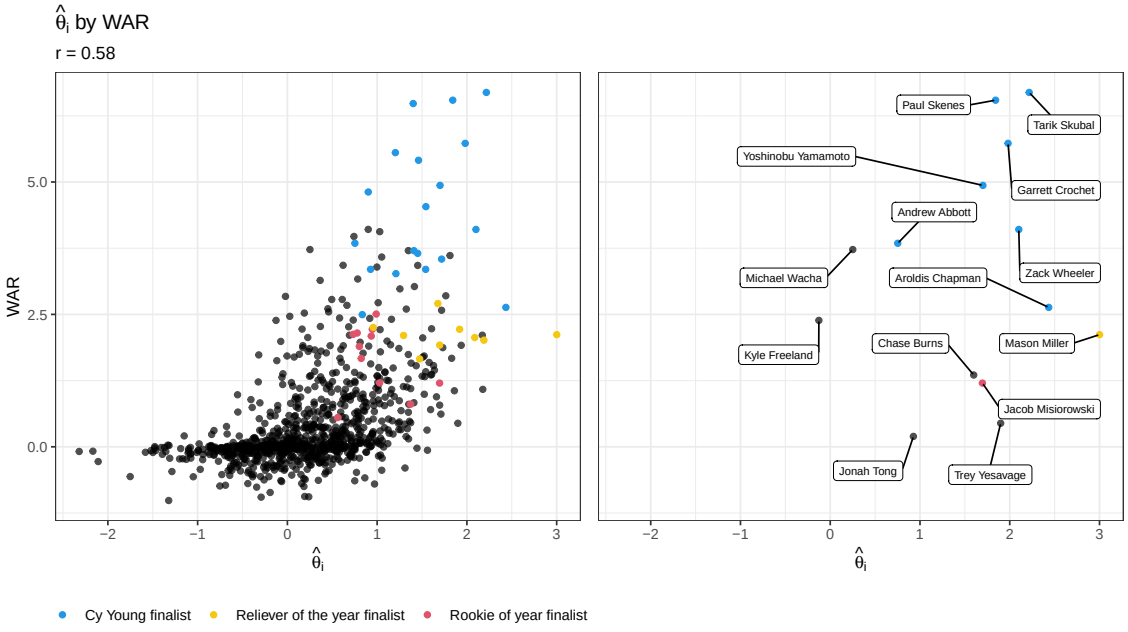

Figure 2: PLATE estimates of pitcher abilities on the x-axis against Wins Above Replacement (WAR) on the y-axis. The left panel presents all 873 pitchers with WAR ratings. The right panel presents a selection relevant players.

PLATE identifies elite pitching talent faster than WAR by linking major and minor league ability scales, allowing PLATE to leverage information from pitcher outings in both leagues. The right panel of Figure 2 reveals a surprising finding: Jonah Tong, a top minor league pitching prospect, has a higher estimated pitching ability than Cy Young finalist Andrew Abbott. Tong spent the majority of the 2025 season in Triple A, only pitching in five major league games. While Tong dominated in Triple A, leading the league in strikeouts per nine innings and posting an ERA of *zero*, he had ups and downs during his five major league appearances. In contrast, Abbott pitched 29 games in the majors during the 2025 season. The comparison between Tong and Abbott offers important insights into how PLATE evaluates pitchers: While Abbott was a Cy Young finalist in 2025, he finished second to last in the National League voting. Compared to other Cy Young finalists, Abbott surrendered more hits. Abbott posted an impressive ERA, but a relatively high FIP,

suggesting he benefited from defensive help. Both PLATE and WAR agree that Abbott was less impressive than other elite starting pitchers like Yoshinobu Yamamoto, Garrett Crochet, and Cy Yong winners Tarik Skubal and Paul Skenes. Abbott's PLATE rating of .753 has a percentile rank of .812. While it recognizes that Abbott is a high performing major league pitcher, it also also correctly identifies that Abbott is not playing at the same level as elite league leaders like Skubal, Yamamoto, and Crochet. Tong's PLATE rating of .927 places him at the .87th percentile rank and properly reflects that Tong is a top pitching prospect. However, given Tong's mixed performance in the major leagues, a PLATE rating of .927 appears inappropriately high. An advantage of the Bayesian inferential framework used to fit the PLATE IRTree is that it quantifies uncertainty in its estimates by providing a full posterior distribution for each parameter estimate. When we consider the full posterior, we observe important context that helps interpret Tong's PLATE rating.

Figure 3 presents the full PLATE rating posterior distributions for Abbott and Tong. The mean of Tong's PLATE posterior distribution is higher than Abbott's, but the *width* of his distribution signals considerably more uncertainty. Tong's posterior distribution of PLATE has a high median, reflecting that, as a top pitching prospect, Tong has a high ceiling. However, due to limited appearances and high variance in performance when facing higher strength major league batters, PLATE is highly uncertain about where to place Tong's ability. On the other hand, PLATE is considerably more certain about Abbott's ability, who played 29 games in the major leagues and was remarkably consistent across performances. While the median value of the PLATE posterior is higher for Tong than for Abbott, the 10th percentile of the posterior distribution (an indicator of a pitcher's performance floor) is notably lower for Tong. When using PLATE for pitcher evaluation, we recommend decision makers consider the full posterior distribution, as it is representative of not only player strength, but performance range and consistency. When acquiring minor league players as part of a trade, for example, high posterior means and medians signal a player generally performs at a high level. However, variance in performance and uncertainty in ability is a consideration when making decisions about which pitchers to call up when major league rosters expand on September 1st.[5]

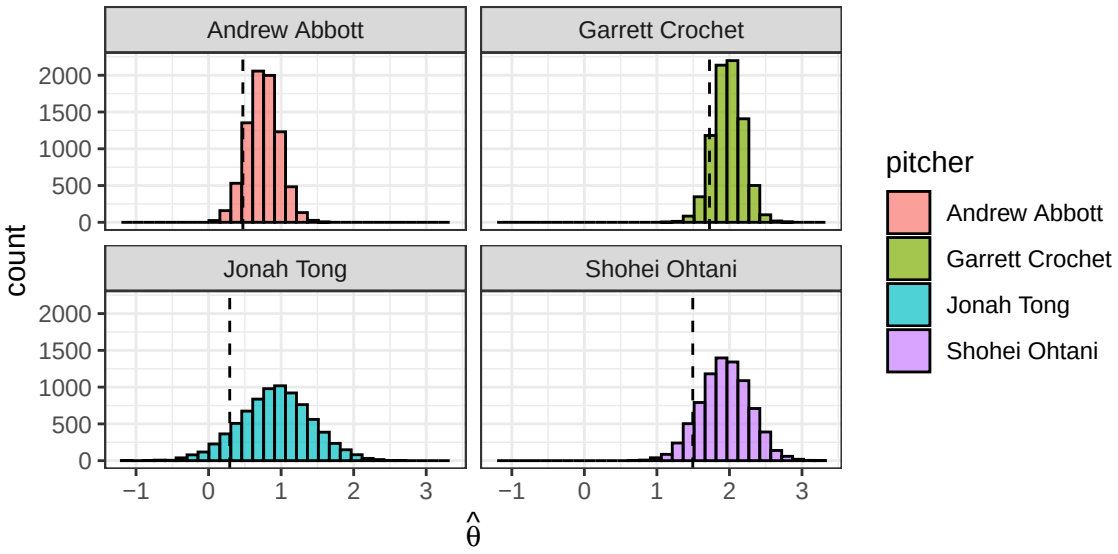

Figure 3: PLATE posterior distributions for Andrew Abbott, Garrett Crochet, Jonah Tong, and Shohei Ohtani. Dashed lines represent the 10th quartile of the posterior distribution.

The uncertainty quantification inherent to PLATE also account for the role of injuries in the assessment of pitcher performance. Both Shohei Ohtani and Garrett Crochet are established elite starting pitchers, but during the 2025 season Ohtani was returning from injury and did not pitch for the first two months of the season as a precautionary measure. While PLATE correctly identifies that Ohtani remained an elite level pitcher, the uncertainty of Ohtani's injury led to a wider posterior distribution when compared to another elite pitcher, like Crochet, who did not face the same injury concerns.

We have shown important discrepancies between WAR and PLATE when evaluating pitchers, but have not considered the role of relief pitching in talent evaluation. Another consequence of weighting by major league innings pitched is that WAR favors starting pitchers over relief pitchers. Mason Miller, arguably the best closer in professional baseball,

[5]On September 1st major league roster sizes can increase from 26 to 28 and additional pitchers can be added to the roster

had the highest PLATE estimate with 3.01 (1.0 percentile rank), greater than Skubal's 2.2 (.99 percentile rank) and Skenes 1.9 (.99 percentile rank). In contrast, Skubal and Skenes have WARs of 6.1 and 6.0, respectively, notably higher than Miller's WAR of 2.5. As a closer, Miller plays fewer innings than starting pitchers Skubal and Skenes, and Miller's limited innings diminish his WAR rating. As such, PLATE and WAR are more similar when considering only starting pitchers who spent the entire season in the major leagues. Restricting to only major league starting pitchers, the Spearman correlation between PLATE and WAR is .79, compared to .58 when we consider all pitchers with WAR ratings.

While WAR favors elite starting pitchers, conventional box score pitching statistics such as ERA, FIP, WHIP, and strikeouts per nine innings (K9) slightly favor relievers and closers, who are often brought in during the final innings of the game, pitching with maximal effort to prevent runs and maintain close leads. Figure 4 presents the Spearman correlations between PLATE and these conventional box statistics. PLATE is also correlated with all statistics in the expected directions. WAR is less correlated with box statistics than PLATE. However, the correlation between PLATE and WAR is comparable to the correlation between PLATE and the conventional pitching statistics presented in Figure 4.

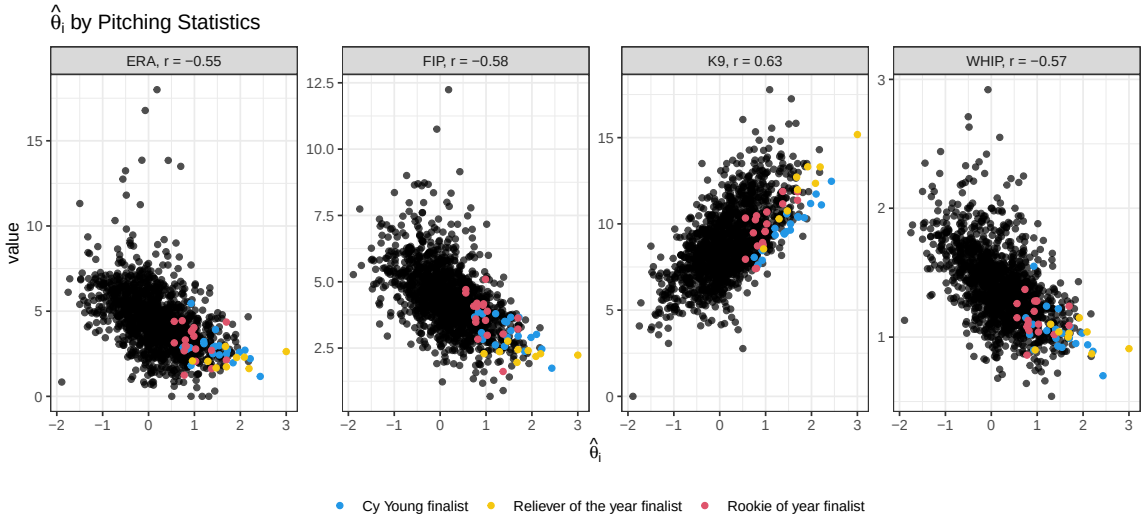

Figure 4: Spearman correlations and bivariate relationships between PLATE and conventional pitching statistics. Unlike WAR which is limited to players who spent time in the major leagues, Figure 4 contains data for both minor league and major league pitching. Pitchers with fewer than 10 innings pitched were removed to account for extreme outliers.

The relationships presented in both Figure 2 and Figure 4 demonstrate the construct validity of PLATE. Unlike the other statistics we have considered so far, a pitcher's PLATE carries over from both the minor and major leagues. To further validate PLATE, we tested the extent to which PLATE can predict a minor league pitcher's performance if they were called up to the major leagues.

For validation purposes, we fit a separate model on all major league and Triple A plate appearances during the 202 season prior to September 1st. On September 1st, major league teams are allowed to expand their rosters from 26 to 28 players, and minor league pitchers are commonly called up to the majors for development or to rest starting pitchers for the playoffs. We used the validation model trained on data before September 1st to predict how well minor league players called up after September 1st would perform in the majors and compared these predictions against pitchers' actual performance.

The absence of a natural outcome is a clear challenge of this validation process. Thus far, we have highlighted the limitations of current methods of pitching evaluation, including WAR, ERA, FIP, and WHIP. Nevertheless, these statistics do capture positive outcomes of pitcher appearances and the ability for PLATE to predict a minor league pitcher's short term ERA, FIP, or WHIP signals the utility of the measure. After training the PLATE validation model on all data prior to September 1st, we forecast ERA, FIP, and WHIP for the 73 minor league players called up after September 1st as a function of the mean and 10th percentile PLATE estimates. To estimate out of sample predictive performance, we report the Leave One Out (LOO) Cross Validated root mean squared error (RMSE). As a reference, we fit separate forecasts using minor league ERA, FIP, and WHIP to predict performance in each metric after being called up to the major leagues. ERA, FIP, and WHIP are on different scales; to compare across metrics, we normalized

the LOO RMSE by dividing by the standard deviation of each metric. After normalization, the LOO RMSE can be interpreted as the average number of standard deviations an out of sample prediction deviates from the true performance. Results of each forecast are presented in Figure 5. Importantly, PLATE-based forecasts for major league ERA, FIP, and WHIP outperform forecasts based on minor league ERA, FIP, and WHIP. For WHIP the difference between the PLATE and minor league WHIP forecasts is small, however, for ERA and FIP the PLATE-based forecasts deliver modest, yet meaningful, improvements in out of sample prediction. Moreover, because PLATE follows a pitcher between leagues, PLATE estimates will rapidly update if a minor league pitcher begins to struggle or dominate in the major leagues. Consider Trey Yesavage: When he was called up, his PLATE rating of .88 (.86 percentile rank) signaled major league potential, but after Yesavage continued elite performance against stronger major league batters, his PLATE rating rose to 1.9 by the end of the season.

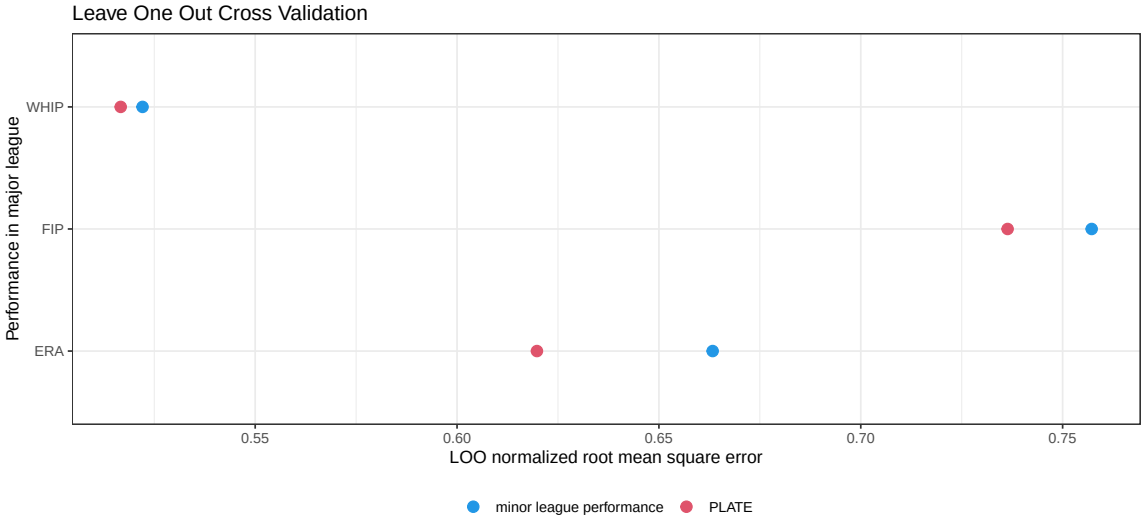

Figure 5: Normalized Leave One Out Cross Validates root mean squared error (RMSE) is shown on the x-axis for corresponding outcomes of major league ERA, FIP, and WHIP, shown on the y-axis. Red points represent the LOO RMSE from a model where the mean and 10th quintile PLATE estimates from minor league pitchers predict major league ERA, FIP, and WHIP. Blue points are comparison models were minor league ERA, FIP, and WHIP are used to predict these outcomes in the major leagues. Models were weighted by games played to account for extreme outliers.

5 Interpretable Batter Parameters

The primary aim of PLATE is to estimate the latent ability of pitchers, but interpretable batter parameters are an ancillary benefit of the estimation process. Figure 6 provides a brief example of the verifying the interpretability of the batter parameters where batter strength for node 1 (strike outs) is plotted against batter strength for node 2 (reaching base given not striking out) for the batters with the top ten strength parameters in node 1 and node 2. Recall that for node 1 high batter strength signals a batter is highly unlikely to strike out. As expected Luis Arraz, who took the fewest strikeouts in the majors, has the highest batter strength in node 1. Likewise Jacob Wilson, who had a breakout rookie year in 2025, only recorded 39 strikeouts over the course of the season.

Within node 2, a higher batter strength indicates a batter is likely to end up on base, conditional on having not struck out. Unsurprisingly, Aaron Judge (who led the major leagues in on base percentage) has the highest batter strength in node 2. Elite batters, such as Shohei Ohtani, and rising stars, such as Roman Anthony, are also recognized with high values in node 2. While the position of Judge and Ohtani is expected, high values for players like Emmanuel Rodriguez may appear surprising. Rodriguez spent all of the 2025 season in Triple A, where he is known for drawing walks and maintains a high on base percentage. Rodriguez is a patient batter who is highly selective about which pitches he will swing at. As a consequence of this style, conditional on not striking out, Rodriguez has a very high probability of getting on base. The high strength parameter assigned to Rodriguez in node 2 reflects these characteristics.

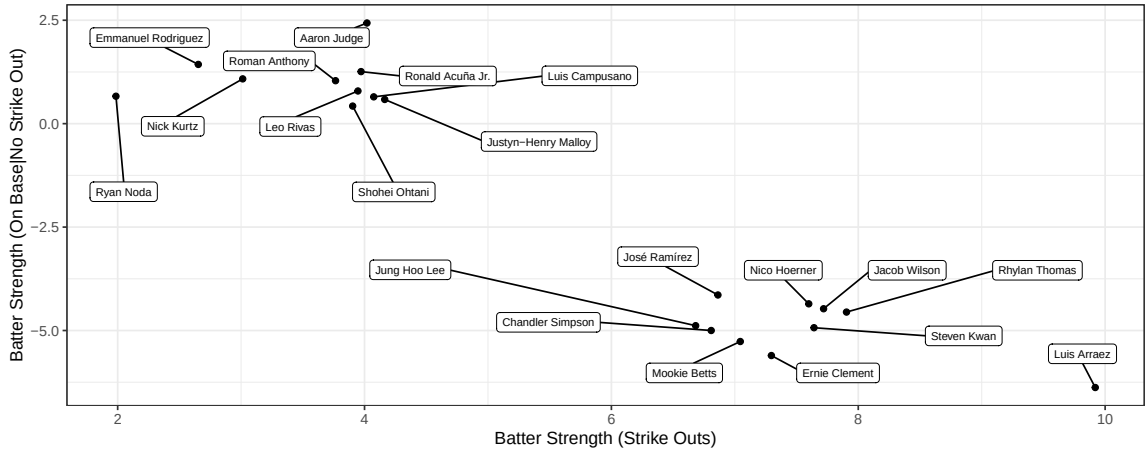

Figure 6: Batter strength parameters for node 1 (x-axis) and batter strength parameters for node 2(y-axis) for the top 10 batters in both nodes 1 and 2. Higher batter strength in node 1 indicates that a batter is *less likely* to strike out. Higher batter strength in node 2 indicates that, conditional on not striking out, a batter is more likely to get on base.

6 Discussion

In this paper, we introduce PLATE, Player Latent Ability Through Events, a novel player evaluation metric based upon Item Response Trees that takes discrete events with categorical outcomes and uses them to estimate player ability on a continuous latent scale. We use pitching ability in professional baseball as our test case, envisioning it as a unidimensional construct and comparing PLATE to established evaluations of pitching performance like WAR, FIP, and ERA. Additionally, we compare PLATE's ability to forecast future performance to these other metrics. Overall, we find that PLATE is better able to identify elite talent, provides meaningful variation at all levels of the ability spectrum, and better predicts future performance than traditional summaries of pitcher skill.

In this context, we see PLATE as delivering three main advantages over traditional pitcher evaluation metrics. First, PLATE accounts for differences in the strengths and styles of batters faced by pitchers, making it more appropriate for players with a limited number of in-game appearances and accounting for variation induced by timing of pitching rotations and team schedule difficulty. This facilitates the second strength of PLATE, in that it puts players from the major and minor leagues onto a common scale, allowing for between-league comparison and better forecasting of future major league performance. Finally, PLATE has inherent uncertainty quantification by providing full posterior distributions on estimated parameters. Considering both PLATE and the width in the posterior distribution of it can facilitate better decision making among management and front office staff, by helping to better understand the relative floor and ceiling on player performance or assess expected levels of in-game consistency. Additionally, comparing the full posterior distributions of PLATE between players can provide more nuanced representations of differences in on field performance.

It is important to understand that PLATE, as presented here, is not the end of the line. Instead, PLATE should be thought of as an extensible and flexible framework for evaluation of player skill from observed events. The power of the IRTree based formulation is that the tree structure can be modified to foreground different aspects of pitcher skill. For example, nodes can be restructured to incorporate next gen stats that value process (like bat speed or ball rotation) instead of just outcome. Different latent variables can be used for pitchers at each node to produce a more nuanced understanding of pitching style. The structure of the tree can even be reversed to produce a one number summary of batter strength and multiple variations on pitcher ability. Finally, while the discrete nature of events in baseball makes it an attractive context for new analytic techniques, there is no reason a IRT-based measurement models such as PLATE are not applicable to other sports or competitive events.

References

[1] Benjamin S Baumer, Shane T Jensen, and Gregory J Matthews. openwar: An open source system for evaluating overall player performance in major league baseball. *Journal of Quantitative Analysis in Sports*, 11(2):69–84, 2015.

[2] Ronald Yurko, Samuel Ventura, and Maksim Horowitz. nflwar: a reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports*, 15(3):163–183, 2019.

[3] Stephen M Shea and Christopher E Baker. Calculating wins over replacement player (worp) for nhl goaltenders. *Journal of Quantitative Analysis in Sports*, 8(1):0, 2012.

[4] Bill Petti and Saiem Gilani. *baseballr: Acquiring and Analyzing Baseball Data*, 2026. R package version 2.0.0.

[5] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025.

[6] Frederic M Lord and Melvin R Novick. *Statistical theories of mental test scores*. Addison Wesley, 1968.

[7] Paul De Boeck and Ivailo Partchev. Irtrees: Tree-based item response models of the glmm family. *Journal of Statistical Software*, 48:1–28, 2012.

[8] Minjeong Jeon and Paul De Boeck. A generalized item response tree model for psychological assessments. *Behavior research methods*, 48(3):1070–1085, 2016.

[9] Fumiko Samejima. Graded response model. In *Handbook of modern item response theory*, pages 85–100. Springer, 1997.

[10] Jiqiang Guo, Jonah Gabry, Ben Goodrich, and S Weber. Package 'rstan'. *URL https://cranr-projectorg/web/packages/rstan2020*, 2020.

[11] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76:1–32, 2017.