A new framework to estimate return on investment for player salaries in the National Basketball Association

Jackson P. Lautier*

May 28, 2024

Abstract

An essential component of financial analysis is a comparison of realized returns. These calculations are straightforward when all cash flows have dollar values. Complexities arise if some flows are nonmonetary, however, such as on court basketball activities. To our knowledge, this problem remains open. We thus present the first known framework to estimate a return on investment for player salaries in the National Basketball Association (NBA). It is a flexible five-part procedure that includes a novel player credit estimator, the Wealth Redistribution Merit Share (WRMS). The WRMS is a per-game wealth redistribution estimator that allocates fractional performance-based credit to players standardized and centered to uniformity. We show it is asymptotically unbiased to the natural share and simultaneously more robust. The per-game approach allows for break-even analysis between high-performing players with frequent missed games and average-performing players with consistent availability. The WRMS may be used to allocate revenue from a single game to each of its players. Using a player's salary as an initial investment, this creates a sequence of cash flows that may be evaluated using traditional financial analysis. We illustrate all methods with empirical estimates from the 2022-2023 NBA regular season. All data and replication code are made available.

Keywords: internal rate of return, load management, player evaluation, player tracking data, sports analytics

1 Introduction

Methods to assess the ongoing financial performance of invested monies are essential for financial analysts. Examples are ubiquitous: mutual fund fact sheets report historical returns, publicly-traded companies report quarterly earnings to shareholders, and lenders report on

*Department of Mathematical Sciences, Bentley University, Waltham, MA
Correspondence: jlautier@bentley.edu

defaulted and delinquent loans. In the vast majority of these cases, both the cash inflows and outflows of invested capital may be recorded as market prices. This makes the financial return calculations rudimentary.

For example, to calculate the realized return on investment (ROI) for a sequence of cash flows, it is possible to utilize the *internal rate of return* (IRR) methodology of Berk and Demarzo (2007, §4.8). That is, we solve for the rate of return, r, such that the discounted present value of future return cash flows equals the time zero investment. Formally, let CF_0 be the initial (i.e., negative) investment, and $\mathrm{CF}_1, \ldots, \mathrm{CF}_K$ be the positive future cash flows. For simplicity, we assume all cash flows occur on equally spaced intervals. Because we are performing a realized, ex post, return calculation, all CF_t, $t = 1, \ldots K$, are assumed known. Then,

$$\left\{ r : \mathrm{CF}_0 = \sum_{t=1}^{K} \frac{\mathrm{CF}_t}{(1+r)^t} \right\} \tag{1}$$

is the realized ROI. Aside from simple forms of (1), solving for r will typically require the use of optimization software (e.g., Varma, 2021).

Complexities arise when one side of (1) does not have a clear monetary cash value or market price, however. One such case is the player contract in the National Basketball Association (NBA). Specifically, given a financial investment into an NBA player via a contractual salary, it is of interest to assess the realized return vis-à-vis on court activities (i.e., points, rebounds, etc.). It is not immediately clear how to value such on court performance in financial terms, and it is this curiosity that is the object of our study. In other words, we endeavor to propose a methodology capable of combining a player's salary and on court performance in such a way as to produce an equivalent formulation of (1). In doing so, we may then solve for r, which is the ROI we desire to estimate.

Financially quantifying on court performance would benefit numerous NBA stakeholders: e.g., informing player evaluations, informing roster building decisions, assessing team roster building competency, and comparing the relative financial efficiency of NBA teams and players. Furthermore, with the recent value of NBA franchises reaching $4 billion (Wojnarowski, 2022), the answers to these questions have become more important than ever. It is natural, then, to suppose there exists a great number of studies that consider both on court performance and salary simultaneously to arrive at methods to measure realized ROI or IRR of a player's contract in view of said player's on court performance. A survey of related studies (e.g., Idson and Kahane, 2000; Berri et al., 2005; Tunaru et al., 2005; Berri and Krautmann, 2006; Berri et al., 2007a; Simmons and Berri, 2011; Halevy et al., 2012; Kuehn, 2017) indicates that this is not the case, however.

We thus propose the first known unified framework to consider both on court performance and salary concomitantly to derive a realized contractual ROI for players in the NBA. It is a five-part process. The first step is to select a measurement period, such as a single NBA regular season. Step two is to select a model to assign fractional credit to players within a single game for all completed games in the measurement time period. Step three is to estimate a Single Game Value (SGV) in dollars for all completed games in the measurement time period. Steps two and three may occur simultaneously after step one. The fourth step is to combine the results of steps two and three to derive player cash flows that are based on relative on court performance. The final step is to use a player's contractual salary as an invested cash flow and the now derived performance-based cash flows to solve for the ROI along the lines of (1). The complete ROI process is summarized in Figure 1.

We illustrate this proposed framework with a novel player credit estimator, the *Wealth Redistribution Merit Share* (WRMS). It is a general estimator that translates an on court player performance estimate into a standardized fractional share, akin to a wealth redistribution exercise that starts from perfect uniformity and reallocates credit via relative performance. We show the WRMS estimator is asymptotically unbiased to the *natural share*, and it is calibrated to a *replacement player*, often desirable in sports analysis (e.g., Shea and Baker, 2012). As an illustration, we present a novel applied study of player performance using logistic regression for data from the 2022-2023 NBA regular season. The attractiveness of the WRMS is that an analyst is free to choose a player performance estimate, and we present

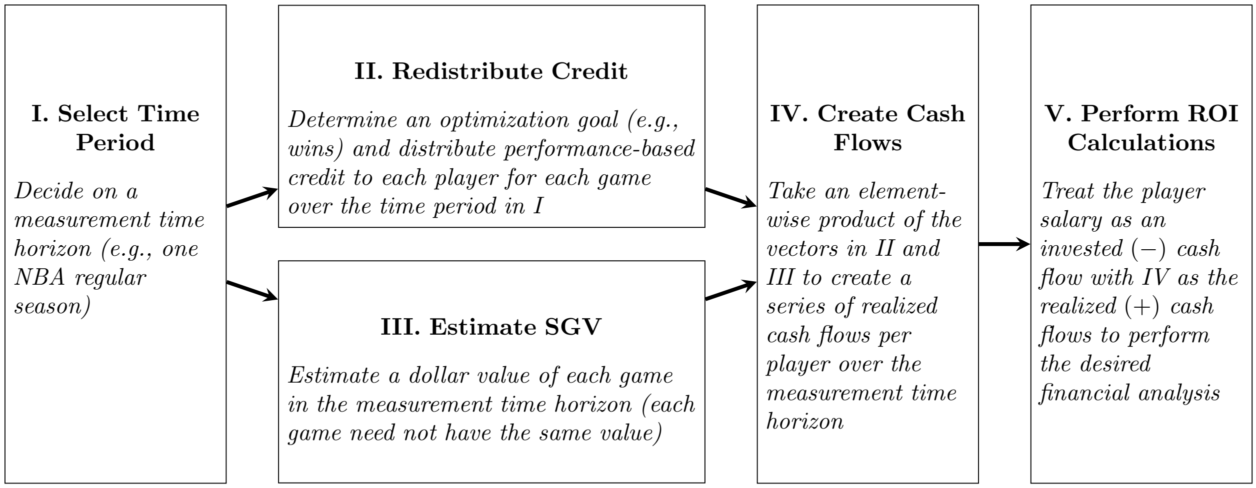

Figure 1: **NBA Contractual ROI Estimation Framework Summary**.

such comparisons. The formal statements of these results may be found in Theorem 2.1. Given we desire to recover (1), our performance measurements are constrained to a single game. This allows us to present a methodology to compare a player with high-performance and frequent missed games against a player with average performance but consistent availability (e.g., Figure 3). To our knowledge, such a perspective remains unexplored in the sports analysis literature. We also propose a model based on ticket sales and television revenue to estimate the SGV. Conditional on the WRMS estimates, Theorem 3.1 ensures our player share dollar estimates are unbiased to total game value.

The paper proceeds as follows. Section 2 begins by heuristically deriving the WRMS starting from the natural share concept and an assumption of complete naivete. Section 2.1 then offers a novel logistic regression player performance measurement, including a review of per-game on court player performance models. The entirety of Section 2 is dedicated to step II in Figure 1. Section 3 then builds upon the work of Section 2 to complete the ROI calculation. It thus includes steps III, IV, and V in Figure 1. In both Sections 2 and 3, we provide empirical illustrations of all methods using data from the 2022-2023 NBA regular season. The paper concludes in Section 4. The Appendix provides complete proofs, and the Supplemental Material includes a brief review of basic finance, a detailed literature review, a glossary of common basketball abbreviations, details on a theoretical derivation of a Cauchy distribution, an index reference, expanded details on the logistic regression model we employ, a comparison of player performance measurements, and simulation studies. All data and replication code used herein may be found at https://github.com/jackson-lautier/nba_roi.

2 Wealth Redistribution Merit Share

The entirety of this section addresses step II of the ROI framework of Figure 1. We first derive the WRMS with a heuristic argument build from the natural share concept. We then expand upon potential on court performance measurement estimators in Section 2.1. Section 2.2 closes with empirical estimates from the 2022-2023 NBA regular season.

To begin, assume there are $N \geq 1, N \in \mathbb{Z}$ total games over the investment horizon selected in step I of Figure 1. Let the current game be denoted by $g \in \mathbb{Z}$, $1 \leq g \leq N$. Per NBA league rules, we assume each team will roster 15 players (National Basketball Association, 2018), and so 30 players within each game have the potential to contribute. We will index each player by $m \in \mathbb{Z}$, $1 \leq m \leq 30$, for each game, g, $1 \leq g \leq N$. It is desirable to only award players that appear in each game (i.e., `MIN` > 0) with credit.[1] This allows us to treat missed games as *defaults* in the ROI framework. In the sequel, we denote the set of players with positive minutes played in game g, $1 \leq g \leq N$, as $\mathcal{M}_g$, and the set of 30 players with the potential to appear in game g, $1 \leq g \leq N$, as $\overline{\mathcal{M}}_g$. Per NBA rules (National Basketball Association, 2018), a minimum of 10 players (5 per team) will receive playing time (i.e., `MIN` > 0). Formally, then, $10 \leq \#\{\mathcal{M}_g\} \leq \#\{\overline{\mathcal{M}}_g\} = 30$ and $\mathcal{M}_g \subset \overline{\mathcal{M}}_g$.

To calibrate the wealth redistribution estimate based upon on court performance, let us first assume there exists some performance measure, $\Delta_{gm} \in \mathbb{R}$, for each player, m, $m \in \overline{\mathcal{M}}_g$, in each game g, $1 \leq g \leq N$. Hence, the *natural player credit game share*, $\mathcal{N}_{gm}$ for player m,

[1] A full glossary of common NBA abbreviations may be found in the Supplemental Material.

$m \in \overline{\mathcal{M}}_g$, in game g, $1 \leq g \leq N$, becomes

$$\mathcal{N}_{gm} = \frac{\Delta_{gm}\mathbf{1}_{m\in\mathcal{M}_g}}{\sum_{\omega\in\overline{\mathcal{M}}_g}\Delta_{g\omega}\mathbf{1}_{\omega\in\mathcal{M}_g}}, \tag{2}$$

where $\mathbf{1}_q = 1$ if statement q is true and 0 otherwise. It is immediate that $\sum_m \mathcal{N}_{gm} = 1$ for all $1 \leq g \leq N$. Intuitively, this implies that players for both teams compete by way of on court performance for a share of the estimated SGV in dollars. Practically, each player m, $m \in \overline{\mathcal{M}}_g$, for game g, $1 \leq g \leq N$, would receive the $\mathcal{N}_{gm}$ percentage share of the SGV. For any player m, $m \in \{\overline{\mathcal{M}}_g \setminus \mathcal{M}_g\}$, $\mathcal{N}_{gm} = 0$ (i.e., players without playing time receive no credit). All subsequent calculations will build from the natural share construct in (2).

As a starting point, we begin with an assumption of complete naivete. Specifically, we assign a degenerative random variable W for Δ_{gm} such that $\Pr(W = c) = 1$, $c \in \mathbb{R}$, for all m, $m \in \overline{\mathcal{M}}_g$, and g, $1 \leq g \leq N$. In this case, the expected credit share of a player $m \in \mathcal{M}_g$, given the total number of players in the set $\mathcal{M}_g$ is known, is the uniform share: the inverse of the cardinality of the set $\mathcal{M}_g$. Symbolically, the uniform credit share is $\mathbf{E}(\mathcal{N}_{gm} \mid \mathcal{M}_g, \Delta_{gm} \sim W) = 1/\#\{\mathcal{M}_g\}$. Hence, we approximate the complete naivete credit share as $1/\mathbf{E}[\#\{\mathcal{M}_g\}]$; that is, the inverse of the average number of players appearing in a game over the measurement time period. If we define $m^* = \sum_g \sum_m \mathbf{1}_{m\in\mathcal{M}_g}$, then an immediate estimator of $1/\mathbf{E}[\#\{\mathcal{M}_g\}]$ is $1/\bar{m}$, where $\bar{m} = m^*/N$.

To incorporate a version of the *replacement player* standardization widely preferred in sports analysis (e.g., Shea and Baker, 2012), we define the sample statistics

$$\bar{\Delta}_{m^*} = \frac{1}{m^*}\sum_{g=1}^{N}\sum_{m\in\mathcal{M}_g}\Delta_{gm}, \tag{3}$$

and

$$s(\Delta_{m^*}) = \sqrt{\frac{1}{m^*-1}\sum_{g=1}^{N}\sum_{m\in\mathcal{M}_g}\left(\Delta_{gm} - \bar{\Delta}_{m^*}\right)^2}. \tag{4}$$

We define *Wealth Redistribution Merit Share* or WRMS as follows.

Theorem 2.1 (Wealth Redistribution Merit Share)**.** *Assume there are $N \geq 1, N \in \mathbb{Z}$, total games over the investment time horizon. Further assume the set $\mathcal{M}_g$ is known for all g, $1 \leq g \leq N$. Let $\mathcal{S} = \{\Delta_{gm}\}_{1 \leq g \leq N, m \in \mathcal{M}_g}$ be a sample of independent and identically distributed (i.i.d.) performance measure random variables. Define the wealth redistribution merit share (WRMS) estimator for player m, $m \in \mathcal{M}_g$ for any game g, $1 \leq g \leq N$, as*

$$\mathcal{W}(\mathcal{S})_{gm} = \frac{1}{s(\Delta_{m^*})}\left(\Delta_{gm} - \bar{\Delta}_{m^*}\right)\frac{1}{\bar{m}} + \frac{1}{\bar{m}}. \tag{5}$$

Then the following three properties hold:

(i) The estimator $\mathcal{W}(\mathcal{S})_{gm}$ is standardized to return a sample mean and sample standard deviation of $1/\bar{m}$ for any $\mathcal{S}$. That is,

$$\frac{1}{m^*}\sum_{g=1}^{N}\sum_{m \in \mathcal{M}_g} \mathcal{W}(\mathcal{S})_{gm} = \sqrt{\frac{1}{m^* - 1}\sum_{g=1}^{N}\sum_{m \in \mathcal{M}_g}\left(\mathcal{W}(\mathcal{S})_{gm} - \frac{1}{\bar{m}}\right)^2} = \frac{1}{\bar{m}}.$$

(ii) For any $\mathcal{S}$, $\mathcal{M}_g$ will be known for all g, $1 \leq g \leq N$. Hence, the bias *of $\mathcal{W}(\mathcal{S})_{gm}$ to the conditional natural share, $\mathcal{N}_{gm} \mid \mathcal{M}_g$, denoted by* $\text{Bias}(\mathcal{W}(\mathcal{S})_{gm}, \mathcal{N}_{gm} \mid \mathcal{M}_g)$*, for all m, $m \in \mathcal{M}_g$, and any g, $1 \leq g \leq N$, is*

$$\text{Bias}(\mathcal{W}(\mathcal{S})_{gm}, \mathcal{N}_{gm} \mid \mathcal{M}_g) = \frac{1}{\bar{m}} - \mathbf{E}(\mathcal{N}_{gm} \mid \mathcal{M}_g) = \frac{1}{\bar{m}} - \frac{1}{\#\{\mathcal{M}_g\}},$$

assuming $\mathbf{E}(\mathcal{N}_{gm} \mid \mathcal{M}_g)$ exists. Further, if $\mathbf{E}(\mathcal{N}_{gm} \mid \mathcal{M}_g)$ exists, then, as $N \to \infty$,

$$\text{Bias}(\mathcal{W}(\mathcal{S})_{gm}, \mathcal{N}_{gm} \mid \mathcal{M}_g) \xrightarrow{p} 0.$$

(iii) Suppose the i.i.d. random variables $\Delta_{gm} \in \mathcal{S}$ are parametric random variables parameterized by $\mathbf{\Theta}$. Let $\hat{\mathbf{\Theta}}_{\text{MLE}} \equiv f(\mathcal{S})$ be a maximum likelihood estimate (MLE) of $\mathbf{\Theta}$. For any function, h_1 of $\mathcal{W}(\mathcal{S})_{gm}$ such that $h_1(\mathcal{W}(\mathcal{S})_{gm}) \equiv h_2(\mathbf{\Theta})$, the maximum likelihood estimate of $h_1(\mathcal{W}(\mathcal{S})_{gm})$ is $h_2(\hat{\mathbf{\Theta}}_{\text{MLE}})$.

Proof. See Appendix A. □

In an economic interpretation, the WRMS of (5) may be thought of as a prescriptive allocation of the SGV share of wealth earned by a player m, $m \in \mathcal{M}_g$, in reference to the performance measure Δ_{gm}, in comparison to uniformity (i.e., complete naivete) for any game g, $1 \leq g \leq N$. Below average games, (i.e., $\Delta_{gm} < \bar{\Delta}_{m^*}$) will decrease the share below $1/\bar{m}$, and above average games (i.e., $\Delta_{gm} > \bar{\Delta}_{m^*}$) will increase the share above $1/\bar{m}$. In effect, then, (5) is a wealth redistribution tool. That is, starting from the complete naivete assumption that all players appearing in a game have equal performance and thus a perfect uniformity of wealth share, the WRMS then redistributes the wealth to each player based on each player's on court performance in comparison to an average (or replacement) player. A notable property of (5) is that players who perform well on the losing team may still receive a large share of the SGV. Finally, observe that by definition

$$\sum_{g=1}^{N} \sum_{m \in \mathcal{M}_g} \mathcal{W}(S)_{gm} = N, \tag{6}$$

which ensures an unbiased estimate at the aggregate level (i.e., the total reallocation of games sums to the original total of games, N).

2.1 Performance Measurement

At present, the i.i.d. on court performance measure random variable, denoted by Δ_{gm} for all m, $m \in \mathcal{M}_g$, and g, $1 \leq g \leq N$, has been left unspecified. Part II of the ROI framework of Figure 1 requires the basketball performance-based calculations to be contained within a single game unit. This is because the overall ROI framework of Figure 1 treats a player's contractual salary as invested capital that is intended to generate per game returns or positive payments. Particularly bad games become negative cash flows (losses), and missed games are treated as *defaults* or missed payments. Outside of the financial ROI framework of Figure 1, the purely basketball importance of the single game unit is well-known (e.g., Oliver, 2004, Chapter 16, pg. 192), and it is thus a natural delineation of NBA performance units. Furthermore, working on a per-game basis offers some advantages. For example, *per possession* standardization (e.g., Oliver, 2004, pg. 25) is not necessary because each team uses approximately the same number of possessions within one game (Berri et al., 2007b, pg. 101). Finally, our per-game approach to performance measurement implies that running season per game totals (e.g., (16) of Section 2.2) allow analysts to determine the exact inflection point of an excellent player that misses many games versus a solid player that consistently plays (e.g., Figure 3.)

Does an existing performance estimator adequately meet our per-game requirements? Given what is available at present, we believe the answer is largely negative. Many previous studies have become dated when compared against recent player tracking data (e.g., Berri, 1999; Page et al., 2007; Fearnhead and Taylor, 2011; Martínez, 2012; Casals and Martínez, 2013). In a promising study, Lackritz and Horowitz (2021) create a model to assign fractional credit to scoring statistics for players in the NBA. Unfortunately, Lackritz and Horowitz (2021) consider only offensive statistics. Idson and Kahane (2000) and Tunaru et al. (2005) do not consider basketball. In a comprehensive review, Terner and Franks (2021) further our findings that a per-game approach is largely unstudied. (The Supplemental Material provides a more detailed literature review.)

One prevalent basketball performance estimator does limit all calculations to a single game: *Game Score* (Sports Reference LLC, 2023). Per (Sports Reference LLC, 2023), Game Score (GmSc) is defined as

$$\begin{aligned} \text{GmSc} = {} & \text{PTS} + 0.4\text{FG} - 0.7\text{FGA} - 0.4(\text{FTA} - \text{FT}) \\ & + 0.7\text{ORB} + 0.3\text{DRB} + \text{STL} + 0.7\text{AST} + 0.7\text{BLK} - 0.4\text{PF} - \text{TOV}, \end{aligned} \tag{7}$$

where the abbreviations follow National Basketball Association (2023).[2] Despite the per-

[2] A full glossary of common NBA abbreviations may be found in the Supplemental Material.

game nature of (7), there are some limitations. First, GmSc does not utilize any of the recent NBA data advancements (National Basketball Association, 2023). Second, it relies on hard-coded coefficients, which are both difficult to interpret without greater context and potentially unstable over time. Finally, GmSc was derived outside of the peer-review process, which has garnered criticism (e.g., Berri and Bradbury, 2010).

There is a much discussed level of subjectivity to assigning credit to players in a basketball game (e.g., Oliver, 2004; Berri et al., 2007b). Given this, it is our intention to propose the general WRMS in Theorem 2.1, of which the analyst is free to choose the performance estimator for Δ. For example, the Win Score (WSc) of Berri et al. (2007b), defined as

$$\begin{aligned}\text{WSc} = \text{PTS} + \text{ORB} + \text{DRB} + \text{STL} + 0.5\text{BLK}\\ + 0.5\text{AST} - \text{FGA} - 0.5\text{FTA} - \text{TOV} - 0.5\text{PF},\end{aligned} \tag{8}$$

may be instead recoded on a per-game basis.[3]

For the purposes of presenting a timely and robust performance measurement model for Δ, we will employ a logistic regression model as follows (Kutner et al., 2005). Let $y_i = 1$ (win) or $y_i = 0$ (loss) with probability $\Pr(y_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{\beta}) \equiv p_i$, where $\boldsymbol{x}_i = (1, X_{i1}, \dots, X_{ik})$ is a row of the design matrix of team level statistics, $\mathbf{X}$. That is, y_i is a Bernoulli random variable with parameter, p_i, for $i = 1, \dots, n$. Notice here the indexing i, $1 \leq i \leq n$ is for game outcome. Hence, for each g, $1 \leq g \leq N = n/2$, there are two game outcomes, $i = 2g$ and $i = 2g - 1$. As we will introduce another indexing variable, j, for the covariates, we provide an index reference in the Supplemental Material.

The formulation of the model implies merit performance credit is directly connected to winning games, though alternative optimization objectives, such as *championships* or *revenue*

[3] A full glossary of common NBA abbreviations may be found in the Supplemental Material.

may instead be used. The binary logit regression model has the form, for $i = 1, \dots, n$,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}. \tag{9}$$

The form (9) implies

$$p_i = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})} = \frac{1}{1+\exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta})}.$$

Hence, the regression coefficients are called log-odds ratios. That is, β_j is the additive increase in the log-odds success probability from a unit increase in x_{ij}, when all other x_{ij^*}'s, $j^* \neq j$, are held fixed, $j, j^* = 1, \dots, k$. Thus, at the team level, any field in $\mathbf{X}$ that returns a positive (and significant) coefficient estimate can be interpreted as having a positive contribution to winning and vice versa for negative coefficients.

Logistic regression in the context of basketball game data outcome offers some pleasing interpretations. First, if we center each covariate, X_{ij}, i.e., replace X_{ij} with $(X_{ij} - \bar{X}_j)$, where $\bar{X}_j = \sum X_{ij}/n$, then the intercept, β_0, becomes the logit at the mean. In other words, an average game by a team yields a $p(\bar{X}_1, \dots, \bar{X}_k) = \exp(\beta_0)/(1+\exp(\beta_0))$ probability of winning. Hence, $\beta_0 = 0$ implies $p(\bar{X}_1, \dots, \bar{X}_k) = 0.5$, a quite reasonable assumption. Second, if we both assume $\beta_0 = 0$ and that each NBA team has the required roster of 15 players per game (National Basketball Association, 2018), then we may distribute the logit of the team's win probability linearly to the logit of each player's individual win probability. This is a direct result of team level statistics equaling the sum of individual player level statistics (with minor exceptions; e.g., a team turnover is not credited to an individual player). We formalize this property in Theorem 2.2.

Theorem 2.2. *Let X_{ijm} represent the individual total for player m, $m = 1, \dots, 15$, for statistical category $j = 1, \dots, k$ for game outcome i, $i = 1, \dots, n$. Fix $j = 1, \dots, k$ and define the team total statistics for game outcome i, $i = 1, \dots, n$, as*

$$X_{ij\bullet} = \sum_{m=1}^{15} X_{ijm}.$$

Then

$$X_{ij\bullet} - \bar{X}_{ij\bullet} = \sum_{m=1}^{15} \Big(X_{ijm} - \bar{X}_{ijm} \Big), \tag{10}$$

where $\bar{X}_{ij\bullet} = \sum_i X_{ij\bullet}/n$ *and* $\bar{X}_{ijm} = \sum_i \sum_m X_{ijm}/15n$*. Further, if we assume* $\beta_0 = 0$ *and recall* (9)*, then*

$$\text{logit}(p_i) = (\boldsymbol{x}_i^*)^\top \boldsymbol{\beta} = \sum_{m=1}^{15} \boldsymbol{x}_{im}^\top \boldsymbol{\beta} = \sum_{m=1}^{15} \text{logit}(p_{im}), \tag{11}$$

where p_i *is the win probability for game outcome* i*,* $i = 1, \ldots, n$*,* $(\boldsymbol{x}_i^*)^\top = (X_{i1\bullet} - \bar{X}_{i1\bullet}, \ldots, X_{ik\bullet} - \bar{X}_{ik\bullet})^\top$*,* $\boldsymbol{x}_{im}^\top = (X_{i1m} - \bar{X}_{i1m}, \ldots, X_{ikm} - \bar{X}_{ikm})^\top$*, and* p_{im} *is the win probability for player* m*,* $m = 1, \ldots, 15$*,*

$$p_{im} = \frac{\exp(\boldsymbol{x}_{im}^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_{im}^\top \boldsymbol{\beta})}.$$

That is, the team level logit of the win probability may be written as a sum of the logits of the individual player win probabilities.

Proof. See Appendix A. □

The first part of Theorem 2.2 may be reminiscent of finding the treatment effects of balanced experiment designs (e.g., Montgomery, 2020).

Remark. *There is an importance assumption of independence underlying the logistic regression model of* (9) *and Theorem 2.2. This independence assumption also plays an important role in Theorem 2.1. For a greater discussion, see Section 4.*

Remark. *We acknowledge an abuse of notation in the indices appearing in Theorem 2.2. Specifically, when the vector notation appears, we drop the* j *covariate index and shift the player index,* m*, to the* j*th position, e.g.,* (11)*. The player index,* m*, also shifts from game,* $1 \leq m \leq 30$*, to team,* $1 \leq m \leq 15$*. We may equivalently index over* $\overline{\mathcal{M}}$ *or* $\mathcal{M}$ *by name,* π*, or* m*,* $1 \leq m \leq 30$*, for any game* g*,* $1 \leq g \leq N$*. This is done beginning at the end of Section 2.2, i.e.,* (15)*. For an index reference, see the Supplemental Material.*

To translate (11) to the performance measurement, Δ_{gm}, $m \in \mathcal{M}_g$, it is necessary to shift the index from game outcome, i, $1 \leq i \leq n$, to game, g, $g = 1, \ldots, n/2$ (recall $N = n/2$).

Hence, to use (11) with Theorem 2.1, we obtain the estimator

$$\mathcal{W}(\boldsymbol{X})_{gm} = \frac{1}{s(\text{WL})_{m^*}}\left(\text{logit}(p_{gm}) - \overline{\text{WL}}_{m^*}\right)\frac{1}{\bar{m}} + \frac{1}{\bar{m}}, \tag{12}$$

where $\overline{\text{WL}}_{m^*} = \sum_g \sum_{m\in\mathcal{M}_g} \text{logit}(p_{gm})/m^*$ and $s(\text{WL})^2_{m^*} = \sum_g \sum_{m\in\mathcal{M}_g}(\text{logit}(p_{gm}) - \overline{\text{WL}}_{m^*})^2$ $/(m^* - 1)$. For the sake of performance measurement comparison, we may also use (7) to define the estimator for player m, $m \in \mathcal{M}_g$ in game g, $g = 1, \ldots, n/2$,

$$\text{GmSc}^*_{gm}(\boldsymbol{X}) = \frac{1}{s(\text{GS})_{m^*}}\left(\text{GmSc}_{gm} - \overline{\text{GS}}_{m^*}\right)\frac{1}{\bar{m}} + \frac{1}{\bar{m}}, \tag{13}$$

where $\overline{\text{GS}}_{m^*} = \sum_g \sum_{m\in\mathcal{M}_g} \text{GmSc}_{gm}/m^*$ and $s(\text{GS})^2_{m^*} = \sum_g \sum_{m\in\mathcal{M}_g}(\text{GmSc}_{gm} - \overline{\text{GS}}_{m^*})^2/(m^* - 1)$. Similarly, via (8) we define for player m, $m \in \mathcal{M}_g$ in game g, $g = 1, \ldots, n/2$,

$$\text{WnSc}^*_{gm}(\boldsymbol{X}) = \frac{1}{s(\text{WS})_{m^*}}\left(\text{WnSc}_{gm} - \overline{\text{WS}}_{m^*}\right)\frac{1}{\bar{m}} + \frac{1}{\bar{m}}, \tag{14}$$

where $\overline{\text{WS}}_{m^*} = \sum_g \sum_{m\in\mathcal{M}_g} \text{WnSc}_{gm}/m^*$ and $s(\text{WS})^2_{m^*} = \sum_g \sum_{m\in\mathcal{M}_g}(\text{WnSc}_{gm} - \overline{\text{WS}}_{m^*})^2$ $/(m^* - 1)$. By property (i) of Theorem 2.1, both (13) and (14) remain equivalently standardized to a sample mean and sample standard deviation of $1/\bar{m}$. Hence, we can directly compare wealth allocation differences between (12), (13), and (14) (e.g., Figure 2).

In closing this section, it may be tempting to ask why (2) cannot be used directly if $\Delta_{gm} \equiv \text{logit}(p_{gm})$ for all $m \in \mathcal{M}_g$, and g, $1 \le g \le N$. The trouble is that, under the assumptions of Theorem 2.2, the conditional natural share in this construct, for any given m, $m \in \mathcal{M}_g$, g, $1 \le g \le N$, is

$$\mathcal{N}_{gm} \mid \mathcal{M}_g, \mathbf{X} = \frac{\text{logit}(p_{gm})}{\sum_{\omega\in\mathcal{M}_g} \text{logit}(p_{g\omega})} \overset{\text{approx}}{\sim} \frac{U}{U+V},$$

where $U \sim N(0, \sigma_u^2)$, $V \sim N(0, \sigma_v^2)$, and $U \perp V$. This is because, with some abuse of notation and allowance for heuristics, $\text{logit}(p_{gm}) \equiv (\boldsymbol{x}^*_{gm})^\top \boldsymbol{\beta} \overset{\text{approx}}{\sim} N(0, \sigma^2)$ (recall $\beta_0 = 0$ by assumption and the covariates are centered). Hence, it can be shown that $U/(U+V)$ follows a Cauchy distribution with location parameter $x_0 = 1/a$ and scale parameter $\gamma = \sqrt{a-1}/a$, where $a = (\sigma_v^2+\sigma_u^2)/\sigma_u^2 = \#\{\mathcal{M}_g\}$ (see the Supplemental Material). Therefore, $\mathbf{E}(\mathcal{N}_{gm} \mid \mathcal{M}_g)$ does not exist! (The median is the location parameter, $1/\#\{\mathcal{M}_g\}$.) Thus, without the stabilization of (5), players would be subject to extreme wealth shares, rendering almost all estimates practically useless. This is an additional advantage of the formulation of (5) in that it is robust to the practical use of a logistic regression model for performance measurement, commonly used in the literature (e.g., Teramoto and Cross, 2010; Daly-Grafstein and Bornn, 2019; Terner and Franks, 2021).

2.2 Empirical Results

We now employ the methods of Section 2.2 to NBA player statistics from the 2022-2023 NBA regular season (National Basketball Association, 2023). To compile an updated set of on court performance statistics, we utilize the `python` package `nba_api` (Patel, 2018). Because we require game-by-game statistics, we design a custom game-by-game query wrapper for Patel (2018). The result is a novel data set of 1,226 2022-2023 NBA regular season games (i.e., $n = 2{,}452$) spanning 36 statistical categories (see the Supplemental Material for details). For completeness, we note that four games did not report player tracking data and were excluded: GSW @ SAS on January 13, 2023, CHI @ DET on January 19, 2023, POR @ SAS on April 6, 2023, and MIN @ SAS on April 8, 2023. To obtain the data and replication code, please navigate to the public `github` repository at https://github.com/jackson-lautier/nba_roi.

In constructing the initial logistic regression and selecting the 36 data fields, we employ three modeling principles: aligning merit to winning, valuing as much on court activity as possible, and avoiding double counting. The variable selection process consists of first fitting a logistic regression model at the team level for all 36 statistical on court data fields. For simplicity, we then remove covariates that are not statistically significant at $\alpha = 0.10$ and perform a second logistic regression. In this second model, we estimate $\hat{\beta}_0 = -0.004930$

Field	Coefficient Estimate	Standard Error	Significance	Variable Importance
FG2O	0.251	0.0267	$***$	9.40
FG2X	-0.349	0.0274	$***$	12.73
FG3O	0.537	0.0368	$***$	14.62
FG3X	-0.368	0.0283	$***$	13.01
FTMO	0.122	0.0221	$***$	5.52
FTMX	-0.220	0.0350	$***$	6.31
PF	-0.197	0.0224	$***$	8.76
AORB	0.356	0.0437	$***$	8.15
ADRB	0.316	0.0246	$***$	12.84
STL	0.443	0.0354	$***$	12.52
BLK	0.132	0.0336	$***$	3.92
TOV	-0.347	0.0292	$***$	11.85
PFD	0.214	0.0329	$***$	6.51
SAST	0.076	0.0214	$***$	3.56
CHGD	0.522	0.1008	$***$	5.18
AC2P	0.041	0.0117	$***$	3.48
C3P	-0.067	0.0140	$***$	4.81
DBOX	0.053	0.0242	$*$	2.18
DFGO	-0.230	0.0179	$***$	12.81
DFGX	0.086	0.0133	$***$	6.50
DDIS	-1.000	0.2009	$***$	4.98
APM	0.016	0.0031	$***$	5.25
OCRB	0.290	0.0371	$***$	7.81
DCRB	0.338	0.0338	$***$	9.99

Table 1: **Logistic Regression Model Parameters**. Based on team outcomes for the 2022-2023 NBA regular season. Because player tracking data was not available for four games, n = 2,452. Significant at $\alpha = 0.001$ ($***$), $\alpha = 0.01$ ($**$), and $\alpha = 0.05$ ($*$). The McFadden R^2 (McFadden, 1974) is 0.6457. Variable importance computed using Kuhn (2008).

with a p-value of 0.948. Hence, we may comfortably refit the logistical regression without an intercept, as it only results in a negligible amount of bias. Because we may use Theorem 2.2 with $\beta_0 = 0$, we feel allowing such small estimation bias is a negligible trade-off (further, the form of (12) will correct this bias per (6)). The final fitted model may be found in Table 1. For reference, the Supplemental Material contains additional details of the model fitting process, such as an expanded discussion on the modeling principles, definitions of each of the original 36 data fields, and the original fitted model with all 36 data fields.

The model of Table 1 suggests that missing shots (i.e., FG2X, FG3X, FTMX), committing fouls (PF) and turnovers (TOV), contesting three point shots (C3P), allowing baskets on defended shots (DFGO), and defensive distance traveled (DDIS) negatively impact win probability. Of these, the only surprise is C3P, though it may be highly related to opponents making three point shots. On the winning side, it is beneficial to make baskets (i.e., FG2O, FG3O, FTMO), collect rebounds (AORB, ADRB), steals (STL), blocks (BLK), draw non-charge fouls (PFD), draw charges (CHGD), set screen assists (SAST), contest two-point shots (AC2P), box out on the defensive end (DBOX), have contested shots miss (DFGX), make passes not counted in assists (APM), and collect contested rebounds (OCRB, DCRB). The most important statistical categories may be assessed by a standard variable importance analysis (Kuhn, 2008). It finds that making (FG3O) and missing (FG3X) three-point field goals are the most important determinants of winning. This aligns closely with long-term trend analysis of the NBA (e.g., Goldsberry, 2019).

The performance measurement model in Table 1 is just one possibility for Δ in (5). Many choices exist, such as (7) and (8). Different choices for Δ will impact the resulting wealth redistribution, which allows an analyst to tailor player credit by performance measurement preference. To illustrate this, we compare the resulting distributions of (12), (13), and (14) in Figure 2. We see that despite having the same mean and standard deviation of $1/\bar{m} = 4.75\%$, the distributions differ. Specifically, the WRMS estimate is more symmetric, whereas both the Game Score and Win Score are skewed right. In a robustness analysis, we find (12) outperforms both (13) and (14) in terms of team win prediction and team rank for data from the 2022-2023 NBA regular season (for details, see the Supplemental Material). As such, the remainder of the manuscript will provide results for (12) only, and the Supplemental Material will provide greater discussion on performance measurement comparisons between (12), (13), and (14). We emphasize that it is the framework of Figure 1 we propose, of which the NBA analyst has flexibility to replace Δ as they see fit.

We may also assess the cumulative total performance of a player over the investment period with a financial perspective. Denote $\mathcal{P} = \bigcup_g \overline{\mathcal{M}}_g$ as the set of all players with the potential to contribute over the investment horizon. For a player π, $\pi \in \mathcal{P}$, let $\mathcal{G}_\pi$ represent the set of games for which player π's team appeared (i.e., $\#\{\mathcal{G}_\pi\} = 82$ for a standard NBA

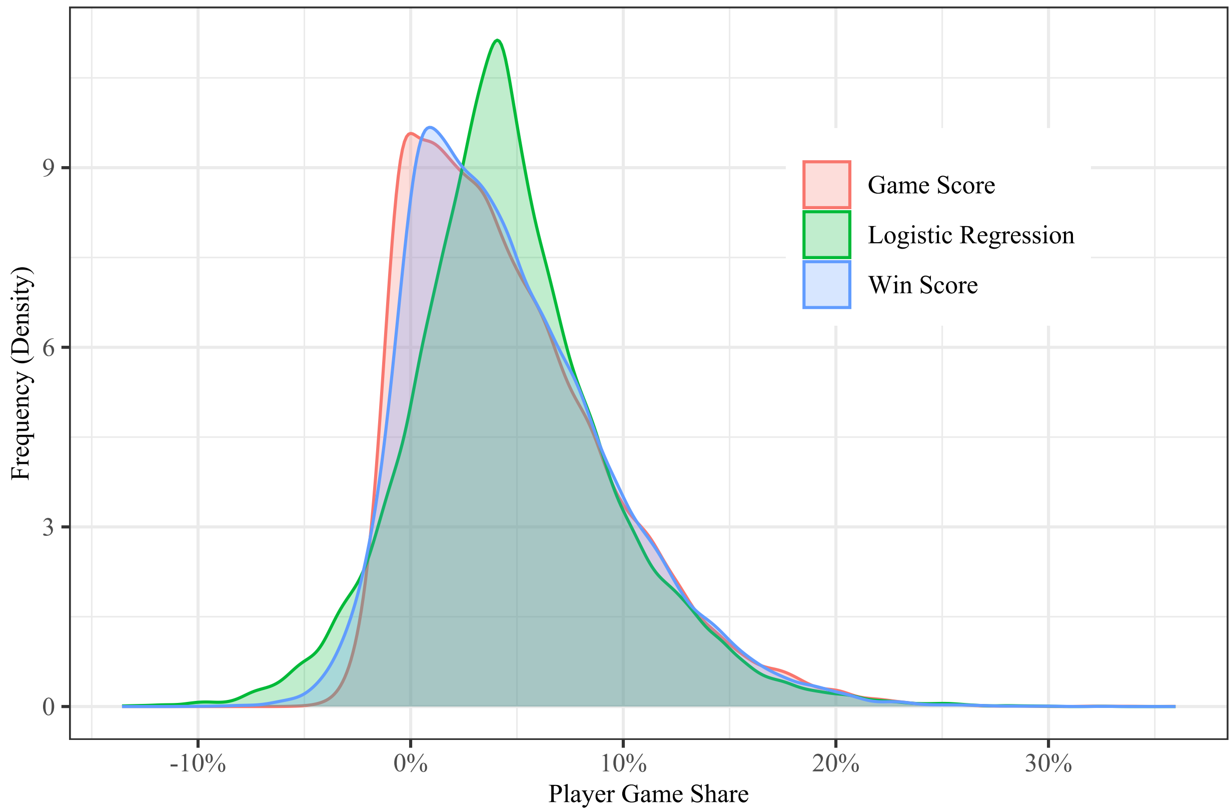

Figure 2: **Wealth Redistribution Comparison**. Frequency distributions of (12), (13), and (14) for all NBA players from the 2022-2023 NBA regular season. The sample of $n = 2{,}452$ game outcomes results in $m^* = 25{,}804$.

regular season). Hence, define for any $g \in \mathcal{G}_\pi$, $\pi \in \mathcal{P}$,

$$\mathcal{W}(\mathcal{S})^*_{g\pi} = \begin{cases} \mathcal{W}(\mathcal{S})_{g\pi}, & \pi \in \mathcal{M}_g \\ 0, & \pi \notin \mathcal{M}_g. \end{cases} \tag{15}$$

Because $\sum_{g=1}^{N} \sum_{m \in \mathcal{M}_g} \mathcal{W}(\mathcal{S})_{gm} = \sum_{g=1}^{N} \sum_{\pi \in \overline{\mathcal{M}}_g} \mathcal{W}(\mathcal{S})^*_{g\pi} = N$ still holds trivially, the desirable unbiased property of (6) remains. In financial parlance, the form of (15) implies a missed game is a *default*. The season total of (15) for player π, $\pi \in \mathcal{P}$, is then

$$\text{PVW}(\cdot)_\pi = \sum_{g \in \mathcal{G}_m} \mathcal{W}(\mathcal{S})^*_{g\pi}. \tag{16}$$

We may consider (16) as a present value of a series of cash flows taking the value of (15) discounted at a zero interest rate. In other words, (16) assumes all single game values are unity. This allows for a pure performance measure that does not include salary. Notably, the game-by-game approach including zeros used in (15) allows for an instant comparison of a high-performing player with frequent missed games against an average-performing player with consistent availability (i.e., Figure 3). This has been a source of perturbation in evaluating players among NBA pundits (e.g., Lowe, 2020), of which (16) may offer new insights.

The placeholder $(\cdot)$ in (16) is generic notation that may be replaced to remind us which performance measurement underlies $\mathcal{W}$. For example, we will use PVWL in the sequel to denote (16) that uses (12) for Δ. For reference, a summary of the distributions of PVWL by position may be found in Figure 4. We can see the model of Table 1 tends to prefer the center position. In addition, we also report the top performing players, of which Nikola Jokic is the top overall PVWL performer. Though outside the scope of our present analysis, we present a comparison of PVW$(\cdot)$ performance measures using (13) and (14) in the Supplemental Material. Because $1/\bar{m} = 4.75\%$, an average player playing 82 games would obtain a PV total of 3.896 for the 2022-2023 NBA regular season, regardless of the performance measure used. For complete results, navigate to the public `github` repository at https://github.com/jackson-lautier/nba_roi.

3 Return on Investment

The purpose of the present section is to complete steps III, IV, and V of the ROI framework of Figure 1. The section proceeds in two parts. First, Section 3.1 introduces a model for the SGV (step III) and an unbiased technique to create the cash flows (step IV). We ultimately reproduce (1) in the NBA context with (19). Section 3.2 then illustrates the ROI framework with data from the 2022-2023 NBA regular season. Prior to this, we briefly review the related literature (the Supplemental Material provides a more detailed literature review).

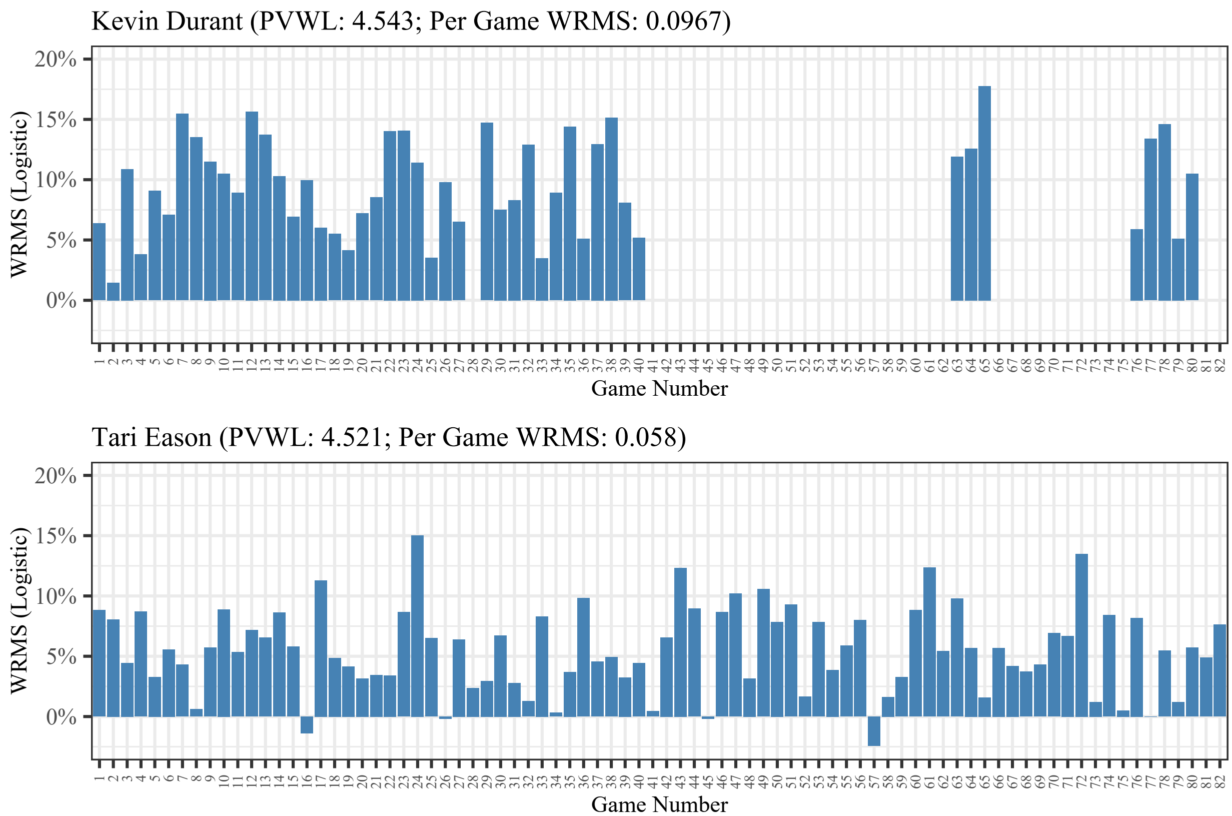

Figure 3: **Quantifying Missed Games**. The per-game approach of (16) allows for break-even calculations between high-performing players with frequent missed games (Kevin Durant, 47 games played, top) against average-performing players with consistent availability (Tari Eason, 82 games played, bottom). Data spans the 2022-2023 NBA regular season.

While no NBA studies consider both player salary and on court performance simultaneously, there is related work outside of basketball (e.g., Idson and Kahane, 2000; Tunaru et al., 2005). The field of sports economics within basketball considers competitive imbalances (Berri et al., 2005), shirking (Berri and Krautmann, 2006), and salaries (Berri et al., 2007a; Simmons and Berri, 2011; Halevy et al., 2012; Kuehn, 2017). Our forthcoming analysis differs from all of these studies generally in that we do not attempt to explain salary decisions. Instead, we propose the first known framework to measure the realized return of a player's contract in light of on court performance.

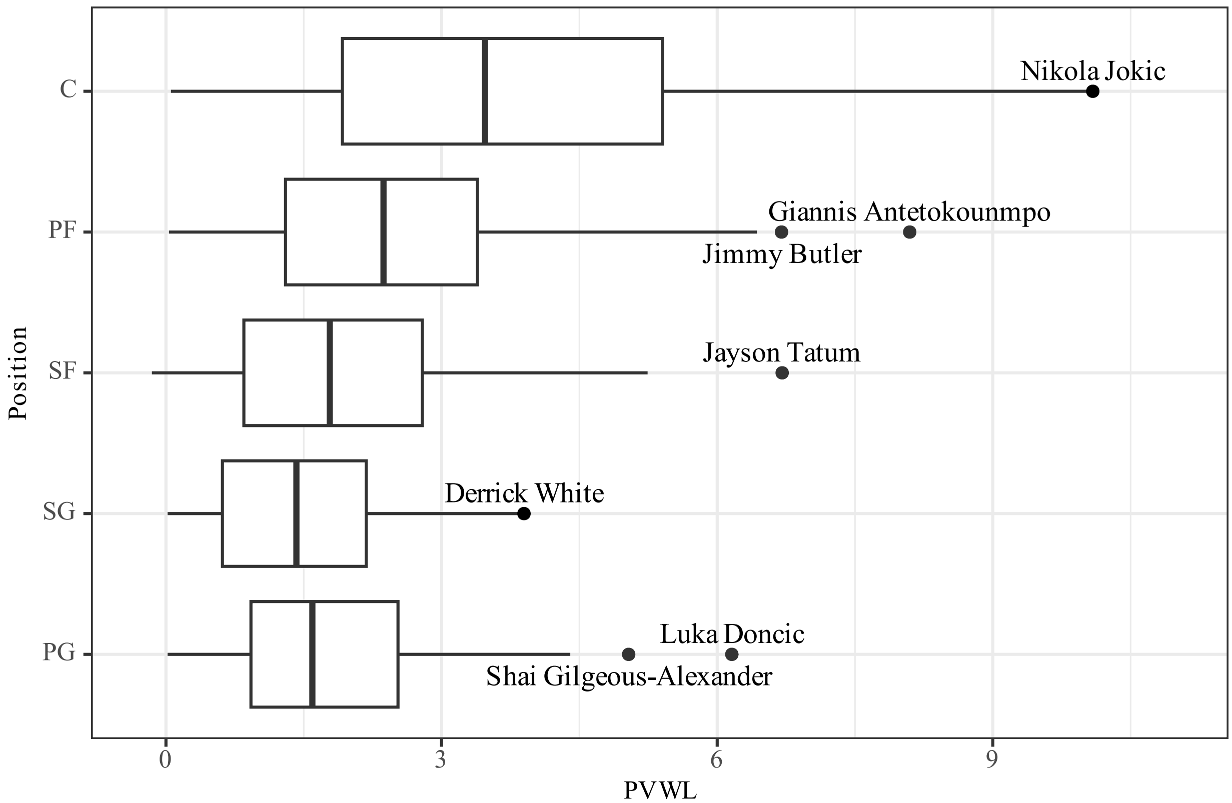

Figure 4: **Top Performers: PVWL**. A summary of the top performers using (16) with logistic regression as the performance measurement (i.e., Table 1) in the WRMS by position. Data spans the 2022-2023 NBA regular season.

3.1 Methods

It remains to estimate the SGV (step III), derive the performance-based cash flows (step IV), and perform the ROI calculations (step V) to complete the ROI framework of Figure 1. Specifically, we first propose a method to model the SGV. Next, we use the SGV model and the results of Section 2.1 to derive an unbiased estimate of a player's performance-based cash flows. Finally, we produce (19) in the form of (1), which results in a player's ROI estimate.

Modeling a SGV is equivalent to answering the question: how does a regular season NBA game generate revenue? Variations of this question have attracted previous attention (e.g., Berri et al., 2007b, Chapter 5). In working from the basic ideas of Berri et al. (2007b), we assume NBA revenue is generated from ticket sales and television rights. We add a third component, which is revenue from advertising. Specifically, for $g = 1, \ldots, N$, define the parametric random variable

$$\mathrm{SGV}_g(\boldsymbol{\alpha}) = \alpha_1 \mathrm{GATE}_g + \alpha_2 \mathbf{1}_{\mathrm{ESPN}} + \alpha_3 \mathbf{1}_{\mathrm{TNT}} + \alpha_4(\mathbf{1}_{\mathrm{ESPN}} + \mathbf{1}_{\mathrm{TNT}} + \mathbf{1}_{\mathrm{NBATV}}), \tag{17}$$

where the parameter vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^\top$ consists of α_1, the average ticket price for an NBA regular season game, α_2, the average TV contract revenue for a regular season NBA game on ESPN, α_3, the average TV contract revenue for a regular season game on TNT, and, α_4, the average advertising revenue for a televised regular season game. Further, GATE_g is a random variable that represents the attendance for game g, $1 \leq g \leq N$. In proposing (17), we do not assume a game televised on NBATV generates television rights revenue for the NBA, but we do assume it generates advertising revenue.

In words, we propose to model SGV_g as the sum total of ticket sales, television revenue, and advertising revenue from game g, $g = 1, \ldots, N$. The natural assumption is that games with higher attendance will be worth more, all else equal, and games that are nationally televised will be worth more, all else equal. This allows us to approximate the relative importance of a game, and it results in the intuitive outcome that players with more nationally televised games will generate a better ROI. This latter point connects with previous studies, as part of the value of signing star players is greater attention from fans and advertisers (e.g., Berri et al., 2007b, Chapter 5).

With an approach to model the SGVs in hand, we may move to deriving the performance-based cash flows (i.e., step IV in Figure 1). In doing so, we will have recovered (1), which is the main objective of our analysis. We first assume the time zero cash flow (i.e., CF_0) is a player's full salary over the investment time horizon and is paid in a single lump sum. For example, assuming an NBA regular season, CF_0 would represent a full season salary. From the perspective of the NBA team, it is a negative cash flow and represents the initial investment. To find the return cash flows, CF_t, $t = 1, \ldots K$, for any player, π, $\pi \in \mathcal{P}$, it is left to multiply (17) with (15) for all $g \in \mathcal{G}_\pi$. This product is player π's, $\pi \in \mathcal{P}$, dollar share of SGV_g, $1 \leq g \leq N$, based on player π's, $\pi \in \mathcal{P}$, on court performance.

Formally, for any player, π, $\pi \in \mathcal{P}$, let $\mathbf{SGV}_{g\in\mathcal{G}_\pi} = (\mathrm{SGV}_1, \ldots, \mathrm{SGV}_K)^\top$ be a vector of SGVs, via (17), and let $\mathbf{W}_{g\in\mathcal{G}_\pi} = (\mathcal{W}^*_{1\pi}, \ldots, \mathcal{W}^*_{K\pi})^\top$ be a vector of WRMSs, via (15), for all games in which player π's, $\pi \in \mathcal{P}$, team appeared over the investment time horizon, where $\#\{\mathcal{G}_\pi\} = K \in \mathbb{N}$. Then the vector of return cash flows over the investment time horizon for player π, $\pi \in \mathcal{P}$, becomes

$$\mathbf{CF}_\pi = (\mathbf{SGV}_{g\in\mathcal{G}_\pi})^\top \mathrm{diag}(\mathbf{W}_{g\in\mathcal{G}_\pi}) = (\mathrm{SGV}_1\mathcal{W}^*_{1\pi}, \ldots, \mathrm{SGV}_K\mathcal{W}^*_{K\pi})^\top, \tag{18}$$

where $\mathrm{diag}(\mathbf{W}_{g\in\mathcal{G}_\pi})$ represents a diagonal $K \times K$ matrix with diagonal $\mathbf{W}_{g\in\mathcal{G}_\pi}$. By the definition of (5), it is possible a particularly bad game may result in $\mathrm{SGV}_t\mathcal{W}^*_{t\pi} < 0$ for some t, $t = 1, \ldots, K$ and player π, $\pi \in \mathcal{P}$.

Before proceeding to complete the ROI methodology, we illustrate that the form (18) has a desirable conditional unbiasedness property. Specifically, recall that (5) may be thought of as a wealth redistribution model that reallocates the SGV based on a player's on court performance. Hence, it is of interest to ensure the reallocated cash flows in (18), given a performance model in (5), are unbiased to the expected sum total of all SGVs, i.e., $\mathbf{E}(\sum_g \mathrm{SGV}_g)$. In other words, we do not wish to inadvertently "create" or "eliminate" wealth due to a faulty estimator. This property holds if $\mathbf{E}(\mathrm{SGV}_g) = \mu \in \mathbb{R}$ for all $g = 1, \ldots, N$.

Theorem 3.1. *Let SGV_g be a single game value random variable for any game, $g = 1, \ldots, N$ such that $\mathbf{E}(\mathrm{SGV}_g) = \mu \in \mathbb{R}$ for all $g = 1, \ldots, N$. Then, conditional on $\mathcal{W}^*_{g\pi}$ for all π, $\pi \in \mathcal{P}$, $g = 1, \ldots, N$,*

$$\mathbf{E}\left(\sum_{g=1}^{N} \sum_{\pi \in \overline{\mathcal{M}}_g} \mathrm{SGV}_g \mathcal{W}^*_{g\pi} \middle| \mathcal{W}^*_{g\pi} \right) = \mu N.$$

That is, the WRMS estimator of (5), when viewed over all players and games in the investment time horizon, is unbiased to the expected total generated revenue.

Proof. See Appendix A. □

Finally, to retrieve the form of (1), let $\boldsymbol{\nu}_\pi = ((1+r_\pi)^{-1}, \ldots, (1+r_\pi)^{-K})^\top$ be a vector of discount factors at the rate, r_π, where $\pi \in \mathcal{P}$. Then the contractual ROI for player π, $\pi \in \mathcal{P}$, over the investment time horizon, is the rate, r_π, that equates the discounted present value of player π's, $\pi \in \mathcal{P}$, cash flows, (18), to player π's, $\pi \in \mathcal{P}$, salary. That is,

$$\left\{ r_\pi : \mathrm{CF}_0^\pi = (\mathbf{SGV}_{g \in \mathcal{G}_\pi})^\top \mathrm{diag}(\mathbf{W}_{g \in \mathcal{G}_\pi}) \boldsymbol{\nu}_\pi \equiv \sum_{t=1}^{K} \frac{\mathrm{SGV}_t \mathcal{W}_{t\pi}^*}{(1+r_\pi)^t} \right\}, \tag{19}$$

where CF_0^π is player π's, $\pi \in \mathcal{P}$, full salary over the investment time horizon. We have thus recovered (1), which completes the ROI framework of Figure 1. We remark that (19) relies on a set of reasonable assumptions, which are discussed more fully in Section 4.

3.2 Empirical Results

We now employ the methods of Section 3.1 to estimate the ROI for player salaries for the 2022-2023 NBA regular season. Player salary data for all players from the 2022-2023 NBA regular season are via HoopsHype (2023) (with one supplement for the player Chance Comanche (Spotrac, 2023)). The data to estimate the parameters of the SGV, denoted by (17), may be compiled from various publicly available sources. As we review the parameter estimates of (17), we will detail these sources. To obtain the data and replication code, please navigate to the public `github` repository at https://github.com/jackson-lautier/nba_roi.

Let us first estimate the parameters of (17) before proceeding to the ROI calculations. Attendance figures are readily available per game (e.g., National Basketball Association, 2023), which allows for a reliable estimate of GATE_g, $g = 1, \ldots, N$. To estimate α_1, we may work backwards from total NBA revenue. Specifically, total gates for the 2022-2023 NBA regular season are known to be 21.57% of total NBA revenue (Statista, 2023a). Further, total NBA revenue for the 2022-2023 NBA regular season is known to be \$10.58B (Statista, 2023c). Hence, we may estimate total gate revenue at \$10.58 $\times$ 21.57% = \$2.28B. With total attendance for the 2022-2023 NBA regular season at 22,234,502 (National Basketball

Coefficient	Description	Estimate
α_1	Ticket Price	\$102.64
α_2	ESPN TV Revenue	\$13,861,386
α_3	TNT TV Revenue	\$18,461,538
α_4	Advertising Revenue	\$6,080,586

Table 2: **Component Estimates of $\mathbf{SGV}_g$.** Coefficient estimates of (17) based on available data for the 2022-2023 NBA regular season (National Basketball Association, 2023; Statista, 2023a,c; Lewis, 2023; Statista, 2023b).

Association, 2023), we arrive at an estimate of the average per-ticket price, $\hat{\alpha}_1 = \$102.64$.

To estimate α_2, α_3, and α_4, we may again work backwards from total NBA revenue. Specifically, it is known that total NBA television revenue for the 2022-2023 NBA regular season is \$1.4B for games televised on ESPN (Lewis, 2023) and \$1.2B for games televised on TNT (Lewis, 2023). With 101 games televised on ESPN (National Basketball Association, 2023) and 65 games televised on TNT, we estimate $\hat{\alpha}_2 = \$13{,}861{,}386$ and $\hat{\alpha}_3 = \$18{,}461{,}538$. Finally, total NBA advertising revenue for the 2022-2023 NBA regular season is known to be \$1.66B (Statista, 2023b). As an approximation, we assume total ad revenue to be spread equally among the 273 nationally televised 2022-2023 NBA regular season games (ESPN: 101; TNT: 65; NBATV: 107) (National Basketball Association, 2023). Hence, we estimate $\hat{\alpha}_4 = \$6{,}080{,}586$. A summary of coefficient estimates for (17) may be found in Table 2. For reference, the top five teams in terms of total SGV for the 2022-2023 NBA regular season are LAL (\$908.3M), GSW (\$885.4M), BOS (\$831.1M), PHX (\$766.3M), and PHI (\$708.5M). Each of these teams play in some of the largest television media markets (Sports Media Watch, 2024), which helps to validate these estimates. Players on these teams will generate higher ROIs because the games are more valuable, all else equal.

To estimate contractual ROI, it is necessary to select a performance measurement random variable for Δ. For consistency with Section 2.2, we will use (12) with the missed game adjustment (15). The only restriction is that a player's salary is at or above the 2022-2023 league minimum, \$1,017,781 (RealGM, L.L.C., 2024). Because we treat missed games as defaults, the minimum game restriction is just one game played. Results for all players in

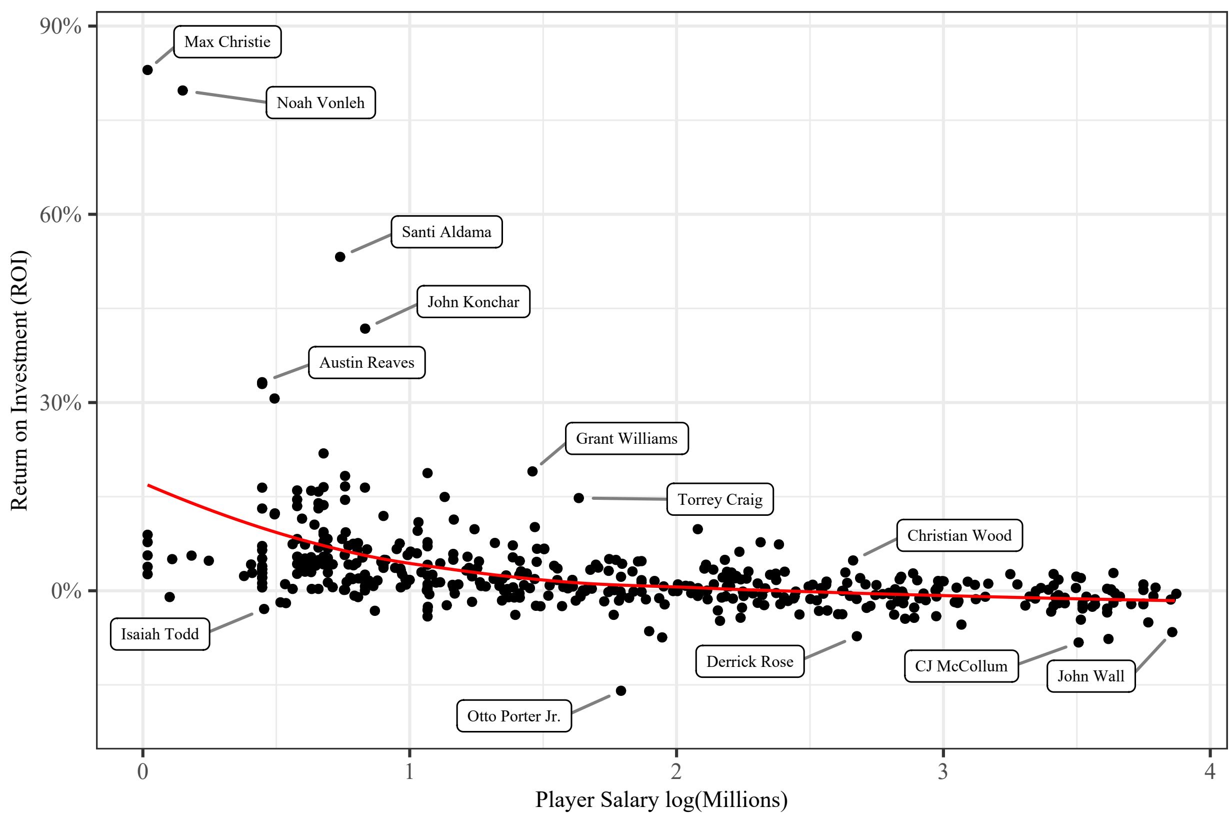

Figure 5: **ROI by Salary: All Players**. A scatter plot of ROI by log of salary for all players with a salary at the league minimum ($1,017,781 (RealGM, L.L.C., 2024)) or higher for the 2022-2023 NBA regular season. The on court performance measurement is (12) with the missed game adjustment (15). Salary data (HoopsHype, 2023; Spotrac, 2023) and SGV parameter estimate data (National Basketball Association, 2023; Statista, 2023a,c; Lewis, 2023; Statista, 2023b; Sports Media Watch, 2024) detailed in Section 3.2. The ROI calculations may be performed using (19).

the 2022-2023 NBA regular season may be found in Figure 5. Not surprisingly, players with higher salaries generally realize lower ROIs, all else equal. The display of Figure 5 may be used by NBA teams to target players that may represent a better relative value at various salary ranges. Similarly, Figure 5 may be used to evaluate the performance of NBA team player personnel decision-makers when signing players. Finally, Figure 5 may be used by the players or player agents in negotiating a new contract that is more closely aligned with comparable players in the aggregate market. To our knowledge, Figure 5 is the first such attempt evaluate the ROI for all players in the NBA.

As an additional illustration of the utility of the ROI estimates of Figure 5, we will use

Position	Coefficient of Variation
Center (C)	2.103
Power Forward (PF)	2.211
Small Forward (SF)	2.940
Shooting Guard (SG)	3.270
Point Guard (PG)	4.710

Table 3: **Coefficient of Variation for ROI by Position**. A ratio of sample standard deviation to sample mean of 2022-2023 NBA regular season empirical ROI estimates in Figure 5 by position.

traditional financial calculations to compare the risk-reward by position. For example, the *coefficient of variation* (CV) (Klugman et al., 2012, Definition 3.2, pg. 20) takes a ratio of the standard deviation of an asset class to its mean. Hence, if we consider each position as an asset class, we may perform the same calculation. We do so in Table 3.

Table 3 suggests that the Center position offers the least risk per unit of return, whereas the Point Guard position is the relative riskiest per unit of return. Such results may be used to help NBA team player personnel decision-makers decide where to invest salary by position, a decision of obvious importance. Furthermore, we may calculate a replacement player ROI. Recall we have normalized (5) to $1/\bar{m}$, which is 4.75% for the 2022-2023 NBA regular season. With an average SGV of \$5,318,785, the combination yields a replacement player game cash flow of \$252,706. Finally, of the 539 players appearing in a 2022-2023 regular season NBA game, we obtain an average salary of \$8,274,410. Therefore, a replacement player appearing in all 82 regular season games yields a 2.71% ROI. As an observation, the ROIs for various players will change with an alternative performance measurement model, such as (13) or (14). For details on this, see the Supplemental Material. For complete results, navigate to the public `github` repository at https://github.com/jackson-lautier/nba_roi.

4 Discussion

A vital component of competently investing in capital markets is assessing the ex post financial performance of invested monies. While such assessments are a standard financial calculation generally, difficulties arise when the returns are non-financial, such as on court basketball activities like rebounding, passing, and scoring. This paper attempts to address these challenges by presenting the first known framework to assess the on court performance of NBA players simultaneously within the relative context of salary. Just as the return on a financial investment is relative to the purchase price, a complete evaluation of player performance is enhanced by considering a player's salary. Such calculations are nontrivial, and the interdisciplinary framework we propose is a five-part process that combines theory from statistics, finance, and economics. With the value of NBA franchises reaching billions of US dollars (Wojnarowski, 2022), the need for such tools is now at an all-time high.

Within the five-part ROI framework we propose in Figure 1, the WRMS of Theorem 2.1 is itself a novel, flexible estimator of player credit capable of considering various estimates of on court player performance. The heuristic derivation of the WRMS suggests a wealth redistribution starting from an assumption of complete naivete. Further, the per-game approach required by (19) yields a new dimension to the field of basketball statistics in the form of break-even calculations for missed games (e.g., Figure 3). Such a calculation is itself timely, as the NBA's governing body has recently implemented strategies to encourage players to avoid missing games (Wimbish, 2023). Pleasingly, the WRMS is asymptotically unbiased to the natural share. To ensure the ROI framework we propose in this manuscript and summarize in Figure 1 is reliable and complete, we use a logistic regression model of player performance. The plug and play design of the ROI framework of Figure 1 allows for analysts to swap out player performance measures, estimators of the SGV, or even the WRMS altogether. It is our intention that this flexibility will be viewed as a positive attribute.

Nonetheless, the infancy of research into methods to combine on court performance with player salaries in the NBA naturally suggests numerous areas ripe for further study. For example, while not necessary to utilize our ROI framework, we elect to constrain our empirical analysis to a single NBA regular season to ease exposition. Player contracts typically span multiple seasons, and so a more complete empirical analysis would increase the observation period. Further, our empirical estimates do not consider play-off games, which some NBA analysts consider to be a significant component of a player's value (Mahoney, 2019). Hence, the empirical ROI estimates may be updated to include the playoffs. Our illustrative logistic regression model in (12) is calibrated to wins, and it is of interest to explore models calibrated to other performance goals, such as championships or revenue. Similarly, the SGV model we propose treats games with higher attendance and viewership as more important. An alternative approach might instead prefer to weight games with a significant impact on the standings as more important (though the two are likely correlated). As an example, Özmen (2016) analyzes the marginal contribution of game statistics across various levels of competitiveness in the Euroleague to win probability. Similarly, Teramoto and Cross (2010) is an example of how weighting schemes may differ for playoff games versus regular season games in the NBA. Something similar may be used to model a game's importance.

An important assumption not yet fully discussed is the implied independence in the sample, $\mathcal{S}$. Though a thorough study is outside the scope of this analysis, discussion is merited. Can players on a basketball court be considered independent? The answer is complex (e.g., Horrace et al., 2022), and more study is needed. For our purposes, the asymptotic unbiasedness derived in Theorem 2.1 will likely maintain if the dependence among the observations is weak enough to allow the Central Limit Theorem to work (Lautier et al., 2023). Hence, as a point estimate, we feel the WRMS concept is likely robust (though we notably do not present any type of variance analysis for this reason). Other approaches, such as mixed effects models or generalized estimating equations could be explored.

The estimators would also benefit from higher precision. This may come through in the form of greater data detail. For example, considering Nielson television ratings, specific ticket prices, or a more refined approach to allocate television revenue. Individual players may get additional credit for off court revenue, such as from jersey sales. A difficulty of these potential enhancements is to obtain detailed data. Higher precision may also be obtained through enhanced calibration. For example, methods exist to refine the quality of a field-goal attempt (e.g., Shortridge et al., 2014; Daly-Grafstein and Bornn, 2019) or account for peer (i.e., teammate) and non-peer effects (e.g., Horrace et al., 2022).

In addition to the statistical aspect, greater precision may be investigated in the financial aspects of the ROI framework of Figure 1 and the derivation of (19). For example, we assume an NBA player's single season salary is paid in one lump sum at time zero. Generally, a player's salary will be paid in installments throughout the regular season. Obtaining more detailed salary payment data will have an impact on the ROI calculations, which may be of interest. Further, we assume all games are played on equally spaced time intervals. This assumption may be explored using financial rate conversion techniques and more precise game dates. Further, an implicit assumption in (19) is that games in the earlier part of the season are given more weight due to the basic conditions of the *time value of money*. Research into the implication of this assumption, such as randomizing the order of the games to calculate a distribution of realized ROI calculations may be prudent. Additionally, the NBA imposes a player salary cap (National Basketball Association, 2018), which includes a team salary floor. Hence, there is an implicit minimum invested, which suggests a type of *risk-free* rate. This may be explored further to offer *Sharpe Ratio* calculations (e.g., Berk and Demarzo, 2007, (11.17)). In addition to the replacement player adjustments employed herein, previous studies such as Niemi (2010) may be helpful for this analysis.

A Proofs

Proof of Theorem 2.1. For the standardization of (i), recall (3), (4), and (5) to write

$$
\begin{aligned}
\frac{1}{m^*}\sum_{g=1}^{N}\sum_{m\in\mathcal{M}_g}\mathcal{W}(\mathcal{S})_{gm} &= \frac{1}{m^*}\sum_{g=1}^{N}\sum_{m\in\mathcal{M}_g}\left(\frac{1}{s(\Delta_{m^*})}\left(\Delta_{gm}-\bar{\Delta}_{m^*}\right)\frac{1}{\bar{m}}+\frac{1}{\bar{m}}\right)\\
&= \frac{1}{\bar{m}}\frac{1}{s(\Delta_{m^*})}\left[\frac{1}{m^*}\sum_{g=1}^{N}\sum_{m\in\mathcal{M}_g}\left(\Delta_{gm}-\bar{\Delta}_{m^*}\right)\right]+\frac{1}{m^*}\sum_{g=1}^{N}\sum_{m\in\mathcal{M}_g}\frac{1}{\bar{m}}\\
&= \frac{1}{\bar{m}}.
\end{aligned}
$$

Next, ignore the radical to similarly show

$$\begin{aligned}\frac{1}{m^*-1}\sum_{g=1}^{N}\sum_{m\in\mathcal{M}_g}\left(\mathcal{W}(\mathcal{S})_{gm}-\frac{1}{\bar{m}}\right)^2 &= \frac{1}{m^*-1}\sum_{g=1}^{N}\sum_{m\in\mathcal{M}_g}\left(\frac{1}{s(\Delta_{m^*})}\left(\Delta_{gm}-\bar{\Delta}_{m^*}\right)\frac{1}{\bar{m}}\right)^2\\ &= \frac{1}{\bar{m}^2}\frac{1}{s(\Delta_{m^*})^2}\frac{1}{m^*-1}\sum_{g=1}^{N}\sum_{m\in\mathcal{M}_g}\left(\Delta_{gm}-\bar{\Delta}_{m^*}\right)^2\\ &= \frac{1}{\bar{m}^2}.\end{aligned}$$

For (ii), recall Δ_{gm} are i.i.d. for all m, $m \in \mathcal{M}_g$, g, $1 \le g \le N$ and observe

$$\begin{aligned}\mathbf{E}(\mathcal{W}(\mathcal{S})_{gm}-\mathcal{N}_{gm}\mid\mathcal{M}_g) &= \mathbf{E}\left(\frac{1}{s(\Delta_{m^*})}\left(\Delta_{gm}-\bar{\Delta}_{m^*}\right)\frac{1}{\bar{m}}+\frac{1}{\bar{m}}-\mathcal{N}_{gm}\mid\mathcal{M}_g\right)\\ &= \frac{1}{\bar{m}}\left(\mathbf{E}\left(\frac{\Delta_{gm}}{s(\Delta_{m^*})}\right)-\frac{1}{m^*}\sum_{g=1}^{N}\sum_{m\in\mathcal{M}_g}\mathbf{E}\left(\frac{\Delta_{gm}}{s(\Delta_{m^*})}\right)\right)+\frac{1}{\bar{m}}\\ &\quad-\mathbf{E}(\mathcal{N}_{gm}\mid\mathcal{M}_g)\\ &= \frac{1}{\bar{m}}-\mathbf{E}(\mathcal{N}_{gm}\mid\mathcal{M}_g).\end{aligned}$$

Further, given $\mathcal{M}_g$, $m \in \mathcal{M}_g$, and g, $1 \le g \le N$,

$$\mathcal{N}_{gm}\mid\mathcal{M}_g = \frac{\Delta_{gm}}{\sum_{\omega\in\mathcal{M}_g}\Delta_{g\omega}}.$$

But Δ_{gm} are i.i.d. for all $\mathcal{S}$, and so the distribution of $\mathcal{N}_{gm} \mid \mathcal{M}_g$ is equivalent for all $m \in \mathcal{M}_g$. Hence, assuming $\mathbf{E}(\mathcal{N}_{gm} \mid \mathcal{M}_g)$ exists,

$$1=\mathbf{E}\left(\frac{\Delta_{g1}+\ldots+\Delta_{g\#\{\mathcal{M}_g\}}}{\Delta_{g1}+\cdots+\Delta_{g\#\{\mathcal{M}_g\}}}\right)=\sum_{m\in\mathcal{M}_g}\mathbf{E}\left(\frac{\Delta_{gm}}{\Delta_{g1}+\cdots+\Delta_{g\#\{\mathcal{M}_g\}}}\right)=\#\{\mathcal{M}_g\}\mathbf{E}(\mathcal{N}_{gm}\mid\mathcal{M}_g),$$

for all $m \in \mathcal{M}_g$. Hence, $\mathbf{E}(\mathcal{N}_{gm} \mid \mathcal{M}_g) = 1/\#\{\mathcal{M}_g\}$. The number of players appearing in any game, g, $1 \le g \le N$, is a discrete random variable over the integers $\{10, \ldots, 30\}$, and so the expectation is finite and nonzero. Hence, by the Weak Law of Large Numbers (Lehmann and Casella, 1998, Theorem 8.2, pg. 54-55) and the continuous mapping theorem (Lehmann and Casella, 1998, Corollary 8.11, pg. 58), consistency follows.

Finally, property (*iii*) is an immediate consequence of the invariance property of the MLE (Mukhopadhyay, 2000, Theorem 7.2.1, pg. 250). □

Proof of Theorem 2.2. Observe,

$$X_{ij\bullet} - \bar{X}_{ij\bullet} = \sum_{m=1}^{15} X_{ijm} - \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{m=1}^{15} X_{ijm}\right) = \sum_{m=1}^{15} X_{ijm} - 15\bar{X}_{ijm} = \sum_{m=1}^{15}\Big(X_{ijm} - \bar{X}_{ijm}\Big).$$

This proves (10). Next, recall (9) with $\boldsymbol{x}_i^\top = (X_{i1\bullet} - \bar{X}_{i1\bullet}, \ldots, X_{ik\bullet} - \bar{X}_{ik\bullet})^\top$ to write via (10)

$$\begin{aligned}
\operatorname{logit}(p_i) = (\boldsymbol{x}_i^*)^\top \boldsymbol{\beta} &= \sum_{j=1}^{k} \beta_j (X_{ij\bullet} - \bar{X}_{ij\bullet}) \\
&= \sum_{j=1}^{k} \beta_j \sum_{m=1}^{15} (X_{ijm} - \bar{X}_{ijm}) \\
&= \sum_{m=1}^{15}\sum_{j=1}^{k} \beta_j (X_{ijm} - \bar{X}_{ijm}) = \sum_{m=1}^{15} \boldsymbol{x}_{im}^\top \boldsymbol{\beta} = \sum_{m=1}^{15} \operatorname{logit}(p_{im}).
\end{aligned}$$

□

Proof of Theorem 3.1. Observe,

$$\begin{aligned}
\mathbf{E}\left(\sum_{g=1}^{N}\sum_{\pi\in\overline{\mathcal{M}}_g} \mathrm{SGV}_g \mathcal{W}_{g\pi}^* \middle| \mathcal{W}_{g\pi}^*\right) &= \sum_{g=1}^{N} \mathbf{E}\left(\sum_{\pi\in\overline{\mathcal{M}}_g} \mathrm{SGV}_g \mathcal{W}_{g\pi}^* \middle| \mathcal{W}_{g\pi}^*\right) \\
&= \sum_{g=1}^{N}\sum_{\pi\in\overline{\mathcal{M}}_g} \mathbf{E}(\mathrm{SGV}_g \mathcal{W}_{g\pi}^* \mid \mathcal{W}_{g\pi}^*) \\
&= \sum_{g=1}^{N}\sum_{\pi\in\overline{\mathcal{M}}_g} \mathbf{E}(\mathrm{SGV}_g) \mathcal{W}_{g\pi}^* \\
&= \mu \sum_{g=1}^{N}\sum_{m\in\mathcal{M}_g} \mathcal{W}_{gm}.
\end{aligned}$$

The proof is then complete by (6). □

References

J. Berk and P. Demarzo (2007). *Corporate Finance, 1st Edition*. Pearson Education, Inc.

D. J. Berri (1999). "Who is 'most valuable'? Measuring the player's production of wins in the National Basketball Association." *Managerial and Decision Economics* **20**, 411–427.

D. J. Berri and J. C. Bradbury (2010). "Working in the land of the metricians." *Journal of Sports Economics* **11**, 29–47.

D. J. Berri, S. L. Brook, B. Frick, A. J. Fenn and R. Vicente-Mayoral (2005). "The short supply of tall people: Competitive imbalance and the National Basketball Association." *Journal of Economic Issues* **39**, 1029–1041.

D. J. Berri, S. L. Brook and M. B. Schmidt (2007a). "Does one simply need to score to score?" *International Journal of Sport Finance* **2**, 190–205.

D. J. Berri and A. C. Krautmann (2006). "Shirking on the court: Testing for the incentive effects of guaranteed pay." *Economic Inquiry* **44**, 536–546.

D. J. Berri, M. B. Short and S. L. Brook (2007b). *The Wages of Wins: Taking Measure of the Many Myths in Modern Sport*. Standford Business Books.

M. Casals and J. A. Martínez (2013). "Modelling player performance in basketball through mixed models." *International Journal of Performance Analysis in Sport* **13**, 64–82.

D. Daly-Grafstein and L. Bornn (2019). "Rao-blackwellizing field goal percentage." *Journal of Quantitative Analysis in Sports* **15**, 85–95.

P. Fearnhead and B. M. Taylor (2011). "On estimating the ability of NBA players." *Journal of Quantitative Analysis in Sports* **7**.

K. Goldsberry (2019). *Sprawlball: A Visual Tour of the New Era of the NBA*. HarperCollins.

N. Halevy, E. Y. Chou, A. D. Galinsky and J. K. Murnighan (2012). "When hierarchy wins: Evidence from the National Basketball Association." *Social Psychological and Personality Science* **3**, 398–406.

HoopsHype (2023). "2022/23 NBA Player Salaries." hoopshype.com. https://hoopshype.com/salaries/players/2022-2023/. Online; accessed 12 June 2023.

W. C. Horrace, H. Jung and S. Sanders (2022). "Network competition and team chemistry in the NBA." *Journal of Business & Economic Statistics* **40**, 35–49.

T. Idson and L. Kahane (2000). "Team effects on compensation: An application to salary determination in the National Hockey League." *Economic Inquiry* **38**, 345–357.

S. A. Klugman, H. H. Panjer and G. E. Willmot (2012). *Loss Models: From Data to Decisions, Fourth Edition*. Hoboken, New Jersey: John Wiley & Sons, Inc.

J. Kuehn (2017). "Accounting for complementary skill sets: Evaluating individual marginal value to a team in the National Basketball Association." *Economic Inquiry* **55**, 1556–1578.

M. Kuhn (2008). "Building predictive models in R using the caret package." *Journal of Statistical Software* **28**, 1–26.

M. H. Kutner, C. J. Nachtsheim, J. Neter and W. Li (2005). *Applied Linear Statistical Models*. McGraw-Hill Irwin.

J. Lackritz and I. Horowitz (2021). "The value of statistics contributing to scoring in the NBA: A quantitative approach." *The American Economist* **66**, 175–189.

J. P. Lautier, V. Pozdnyakov and J. Yan (2023). "Pricing time-to-event contingent cash flows: A discrete-time survival analysis approach." *Insurance: Mathematics and Economics* **110**, 53–71.

E. Lehmann and G. Casella (1998). *Theory of Point Estimation, 2nd Edition*. Springer.

J. Lewis (2023). "NBA announces 9-year extension with ESPN, Turner, through 2025." https://www.sportsmediawatch.com/ - Sports Media Watch https://www.sportsmediawatch.com/2014/10/nba-tv-deal-espn-abc-tnt-nine-year-deal-2025-24-billion-lockout/. Online; accessed 27 December 2023.

Z. Lowe (2020). "2020 NBA awards ballot: Zach Lowe's picks for All-NBA, All-Defensive and All-Rookie." https://www.ESPN.com - Entertainment and Sports Programming Network. https://www.espn.com/nba/story/_/id/29541185/2020-nba-awards-ballot-zach-lowe-picks-all-nba-all-defensive-all-rookie. Online; accessed 24 July 2023.

R. Mahoney (2019). "The 16-game player: Why Draymond Green thrives in the NBA playoffs." Sports Illustrated. https://www.si.com/nba/2019/05/29/draymond-green-warriors-raptors-2019-nba-finals-stephen-curry-kevin-durant. Online; accessed 20 July 2023.

J. A. Martínez (2012). "Factors determining production (FDP) in basketball." *Economics and Business Letters* **1**, 21–29.

D. McFadden (1974). "Conditional logit analysis of qualitative choice behavior." *Frontiers in Econometrics* pp. 105–142.

D. C. Montgomery (2020). *Design and Analysis of Experiments: Tenth Edition*. Hoboken, New Jersey: John Wiley & Sons, Inc.

N. Mukhopadhyay (2000). *Probability and Statistical Inference*. New York, NY: Marcel Dekker.

National Basketball Association (2018). "Highlights of the Collective Bargaining Agreement between the National Basketball Association (NBA) and National Basketball Players Association (NBPA)." nba.com. https://cdn.nba.com/manage/2021/03/2018-19-CBA.pdf. Online; accessed 19 July 2023.

National Basketball Association (2023). "Statistics." nba.com. https://www.nba.com/stats. Online; accessed 13 June 2023.

J. B. Niemi (2010). "Evaluating individual player contributions to team offense and defense: A model based approach." JSM Proceedings, Section on Statistics in Sports, Vancouver, BC, Canada: American Statistical Association.

D. Oliver (2004). *Basketball on Paper: Rules and Tools for Performance Analysis*. Potomac Books, Inc.

M. U. Özmen (2016). "Marginal contribution of game statistics to probability of winning at different levels of competition in basketball: Evidence from the Euroleague." *International Journal of Sports Science & Coaching* **11**, 98–107.

G. L. Page, G. W. Fellingham and C. S. Reese (2007). "Using box-scores to determine a position's contribution to winning basketball games." *Journal of Quantitative Analysis in Sports* **3**.

S. Patel (2018). "nba_api: An API client package to access the APIs of NBA.com." GitHub repository. https://github.com/swar/nba_api. Online; accessed 21 July 2023 (v1.2.0).

RealGM, L.L.C. (2024). "2017 CBA minimum annual salary scale." RealGM.com. https://basketball.realgm.com/nba/info/minimum_scale. Online; accessed 1 May 2024.

S. M. Shea and C. E. Baker (2012). "Calculating Wins over Replacement Player (WORP) for NHL goaltenders." *Journal of Quantitative Analysis in Sports* **8**.

A. Shortridge, K. Goldsberry and M. Adams (2014). "Creating space to shoot: Quantifying spatial relative field goal efficiency in basketball." *Journal of Quantitative Analysis in Sports* **10**, 303–313.

R. Simmons and D. J. Berri (2011). "Mixing the princes and the paupers: Pay and performance in the National Basketball Association." *Labour Economics* **18**, 381–388. The Economics of Sports Labour Markets.

Sports Media Watch (2024). "Major pro and college sports teams ranked by market size." Sports Media Watch. https://www.sportsmediawatch.com/nba-market-size-nfl-mlb-nhl-nielsen-ratings/. Online; accessed 9 February 2024.

Sports Reference LLC (2023). "Glossary." Basketball-Reference.com - Basketball Statistics and History. https://www.basketball-reference.com/about/glossary.html. Online; accessed 26 January 2023.

Spotrac (2023). "Contract details." Spotrac.com. https://www.spotrac.com/nba/portland-trail-blazers/chance-comanche-82139/. Online; accessed 25 July 2023.

Statista (2023a). "Gate receipts as percentage of total revenue in the National Basketball Association from 2010/11 to 2022/23." https://www.statista.com/ - Statista https://www.statista.com/statistics/193410/percentage-of-ticketing-revenue-in-the-nba-since-2006/. Online; accessed 27 December 2023.

Statista (2023b). "National Basketball Association sponsorship revenue from 2010 to 2023." https://www.statista.com/ - Statista https://www.statista.com/statistics/380270/nba-sponsorship-revenue/. Online; accessed 27 December 2023.

Statista (2023c). "National Basketball Association total league revenue from 2001/02 to 2022/23." https://www.statista.com/ - Statista https://www.statista.com/statistics/193467/total-league-revenue-of-the-nba-since-2005/. Online; accessed 27 December 2023.

M. Teramoto and C. L. Cross (2010). "Relative importance of performance factors in winning NBA games in regular season versus playoffs." *Journal of Quantitative Analysis in Sports* **6**.

Z. Terner and A. Franks (2021). "Modeling player and team performance in basketball." *Annual Review of Statistics and Its Application* **8**, 1–23.

R. Tunaru, E. Clark and H. Viney (2005). "An option pricing framework for valuation of football players." *Review of Financial Economics* **14**, 281–295. Real Options SI.

J. Varma (2021). *jrvFinance: Basic finance; NPV/IRR/annuities/bond-Pricing; Black Scholes*. R package version 1.4.3.

J. Wimbish (2023). "NBA CBA: New details emerge about in-season tournament, 65-game threshold for MVP, other awards." https://www.cbssports.com/ - CBS Sports. https://www.cbssports.com/nba/news/nba-cba-new-details-emerge-about-in-season{}-tournament-65-game-threshold-for-mvp-other-awards/. Online; accessed 28 July 2023.

A. Wojnarowski (2022). "Mat Ishbia agrees to Suns purchase for record $4 billion." https://www.ESPN.com - Entertainment and Sports Programming Network. https://www.espn.com/nba/story/_/id/35292815/sources-mat-ishbia-finalizing-suns-purchase-4-billion. Online; accessed 30 June 2023.

NBA ROI: Supplemental Material

The following is intended as an online companion supplement to the manuscript, *A new framework to estimate return on investment for player salaries in the National Basketball Association*. Please attribute any citations to the original manuscript.

This companion includes a brief review of discounting cash flows with interest, a detailed literature review, a glossary of standard statistical abbreviations used in the NBA, a result related to generating a Cauchy distribution, a reference of indexing variables, additional logistic regression model details, and simulation studies (including an extension to Theorem 3.1). Unless otherwise stated, all references are to the main manuscript. All data and replication code is publicly available at the repository: https://github.com/jackson-lautier/nba_roi.

A Financial Review

The objective of the manuscript is to calculate an internal rate of return or realized return on investment for a sequence of cash flows. Such financial parlance may be unfamiliar in statistical circles, and we briefly review the fundamentals here. Let us first review *present value*, which relates to the time value of money. For simplicity, suppose we may earn an annual effective rate of i over the next year. Then, if we owe \$1 one year from today, it is sufficient to invest $\$1/(1+i)$ now because

$$\left(\frac{1}{1+i}\right)(1+i) = 1.$$

As such, financial return calculations routinely consider this time value of money. One example is a sequence of cash flows, which is typically represented in a time line, such as Figure A1. In this case, the future cash flows, CF_t, $t = 1, \ldots, K$, represent realized returns. Conversely, the initial time zero cash flow, CF_0, represents the initial investment. To determine the return, we now seek the rate, r such that the initial investment, CF_0,

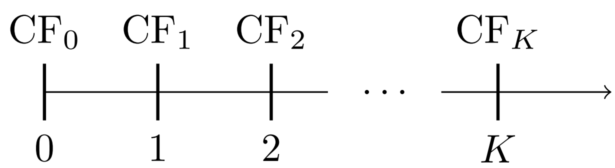

Figure A1: **Cash Flow Time Line**. A classical illustration of a sequence of financial cash flows. The objective of the NBA contractual ROI modeling framework we propose (i.e., Figure 1) is to create a sequence of cash flows in this form, from a combination of salary and on court performance. Once created, it is possible to proceed with standard financial calculations, such as (1).

equals the discounted present value of the future cash flows. This is exactly (1) in Section 1. Many references exist with expanded details, such as Berk and Demarzo (2007).

B Detailed Literature Review

The purpose of this section is to provide more detail to the literature review in the main document, which was abbreviated for ease of exposition. We proceed in two parts. Section B.1 focuses on basketball performance analysis, especially as it relates to the desired properties of the ROI framework of Figure 1. Section B.2 then focuses on financial performance analysis within basketball and sports more generally.

B.1 Performance Measurement

Part II of the ROI framework of Figure 1 requires the basketball performance-based calculations to be contained within a single game unit to better mirror financial analysis. As we find in Section 2.1, a single game performance measurement that also considers more recent player tracking data is not presently available. This motivates the logistic regression analysis we pursue beginning from (9) and expanded upon in Section F. For completeness, we now provide additional detail to the studies referenced in Section 2.1.

Classical regression treatments, such as Berri (1999), do not perform calculations on a game-by-game basis and have become dated considering the advancements in data availability (National Basketball Association, 2023). Data advancements also rule out Page et al. (2007), who fit a hierarchical Bayesian model to 1996-1997 NBA box score data to measure the relative importance of a position to winning basketball games. The same is true for Fearnhead and Taylor (2011), who, in another Bayesian study, propose an NBA player ability assessment model that is calibrated to the relative strength of opponents on the court (via various forms of prior season data; Fearnhead and Taylor (2011) provide results for the 2008-2009 NBA regular season). The work of Casals and Martínez (2013), who fit an OLS model to 2006-2007 NBA regular season data in an attempt to measure the game-to-game variability of a player's contribution to points and Win Score (e.g., Berri et al., 2007b; Berri and Bradbury, 2010), is closer in spirit but does not provide the level of box score detail we desire (the same is true for Martínez (2012)).

B.2 Return on Investment

To our knowledge, no basketball studies consider both player salary and on court performance simultaneously. Per the financial aspects of the ROI framework of Figure 1, we now expand on the related work mentioned only briefly in Section 3.

Idson and Kahane (2000) attempt to derive the determinants of a player's salary in the National Hockey League with a model that incorporates the performance of teammates. We consider the NBA, however, and our methodology differs considerably. Berri et al. (2005) identify the importance of height in the NBA and juxtaposes it against population height distributions to explain competitive imbalances observed in the NBA. Such imbalances are thought to negatively impact economic outcomes of sports leagues (Berri et al., 2005). While financial considerations enter into the analysis of Berri et al. (2005), it does not concern the ROI of single players but rather professional leagues overall. Tunaru et al. (2005) develop a claim contingent framework that is connected to an option style valuation of an on field performance index for football players. Our proposed method differs materially, however, and we focus on basketball rather than football.

Berri and Krautmann (2006) find mixed results to the question of whether or not signing a long-term contract leads to shirking behavior from NBA players. The overall objective of their study differs meaningfully from that of our proposed ROI framework, however. More recently, Simmons and Berri (2011) find salary inequality is effectively independent of player and team performance in the NBA, a result that runs counter to the hypothesis of fairness in traditional labor economics literature. In a related study, Halevy et al. (2012) find the hierarchical structure of pay in the NBA can enhance performance. Neither study attempts to produce a contractual ROI, however. Kuehn (2017) assumes the ultimate goal of each team is to maximize their expected number of wins to find teammates have a significant impact on an individual player's productivity. Kuehn (2017) subsequently reports that player salaries are determined instead mainly by individual offensive production, which can lead to a misalignment of incentives between individual players and team objectives. Of note, the salary findings of Kuehn (2017) correspond to those of Berri et al. (2007a), a similar study.

C Basketball Glossary

The main body of the manuscript assumes some familiarity with the NBA, especially the common statistical abbreviations used in the National Basketball Association (2023). For completeness, we provide a glossary of such abbreviations not defined in the main body of the manuscript (ordered by appearance). All definitions are taken directly from National Basketball Association (2023), which, for reference, also provides a glossary.

MIN (*Minutes Played*) The number of minutes played by a player or team.

PTS (*Points*) The number of points scored.

FG (*Field Goals Made*) The number of field goals that a player or team has made. This includes both 2 pointers and 3 pointers.

FGA (*Field Goals Attempted*) The number of field goals that a player or team has attempted. This includes both 2 pointers and 3 pointers.

FT (*Free Throws Made*) The number of free throws that a player or team has made.

FTA (*Free Throws Attempted*) The number of free throws that a player or team has made.

ORB (*Offensive Rebounds*) The number of rebounds a player or team has collected while they were on offense.

DRB (*Defensive Rebounds*) The number of rebounds a player or team has collected while they were on defense.

STL (*Steals*) Number of times a defensive player or team takes the ball from a player on offense, causing a turnover.

AST (*Assists*) The number of assists – passes that lead directly to a made basket – by a player.

BLK (*Blocks*) A block occurs when an offensive player attempts a shot, and the defense player tips the ball, blocking their chance to score.

PF (*Personal Fouls*) The number of personal fouls a player or team committed.

TOV (*Turnovers*) A turnover occurs when the player or team on offense loses the ball to the defense.

D Cauchy Distribution

The following result is referenced at the close of Section 2.1. Suppose $X \sim N(0, \sigma_x^2)$ and $Y \sim N(0, \sigma_y^2)$, where $X \perp Y$. We show

$$\frac{X}{X+Y} \sim \text{Cauchy}\left(x_0 = \frac{\sigma_x^2}{\sigma_y^2 + \sigma_x^2}, \gamma = \frac{\sigma_y \sigma_x}{\sigma_y^2 + \sigma_x^2} \right). \tag{S.1}$$

Recall,

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{x^2}{2\sigma_x^2} \right) \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{y^2}{2\sigma_y^2} \right).$$

Hence, define $Z = X/(X+Y)$ and $W = X$. By the standard Jacobian transformation (e.g., Mukhopadhyay, 2000, Theorem 4.4.1, pg. 192), the joint probability density function of

(Z, W) is

$$f_{Z,W}(z,w) = \frac{1}{2\pi}\left|\frac{w}{z^2}\right|\frac{1}{\sigma_x\sigma_y}\exp\left(-\frac{w^2}{b}\right),$$

where

$$b = \left(\frac{1}{\sigma_x^2} + \frac{(1-z)^2}{z^2\sigma_y^2}\right)^{-1}.$$

The marginal distribution of Z is then

$$\int_{\mathcal{W}} f_{Z,W}(z,w)dw = \frac{1}{\pi}\frac{b}{\sigma_x\sigma_y z^2}.$$

But,

$$\frac{1}{\pi}\frac{b}{\sigma_x\sigma_y z^2} = \left(\pi\frac{\sqrt{a-1}}{a}\left[1+\left(\frac{z-\frac{1}{a}}{\frac{\sqrt{a-1}}{a}}\right)^2\right]\right)^{-1},$$

where $a = (\sigma_y^2 + \sigma_x^2)/\sigma_x^2$. This is the probability density function of the Cauchy distribution (e.g., Mukhopadhyay, 2000, (1.7.31), pg. 47), which is specified in (S.1). This result may also be confirmed in the simulation studies of Section H.

E A Reference of Indices

The statement of Theorem 2.2 in combination with the WRMS definition of Theorem 2.1 necessitates a series of indexing variables that may be difficult to track. As a reference, we present Table E1. In fitting a logistic regression model to NBA regular season data, there will be n game outcomes. We index each game outcome by i, $1 \leq i \leq n$. Because there are no ties, there will be $n/2 \equiv N$ wins and, similarly, $n/2 \equiv N$ total games. We index each game by g, $1 \leq g \leq N \equiv n/2$, and each game has two game outcomes. Further, we require by Theorem 2.2 that each team roster 15 players for each game. (The roster of 15 is also set by NBA league rules (National Basketball Association, 2018).) This assumption allows us to fit a centered covariate vector at the team level, and then allocate the fitted team level logit to each player depending on the player's individual statistics for game, g,

	Game	Game Outcome	Player			Covariates		
Index	g	i	m			j		
Start	1	1	1			1		
Stop	$N \equiv n/2$	n	15			k		
			1	X_{i11}	$\dots$	X_{ij1}	$\dots$	X_{ik1}
			$\vdots$	$\vdots$	$\ddots$			$\vdots$
		i	m			X_{ijm}		
			$\vdots$	$\vdots$			$\ddots$	$\vdots$
			15	$X_{i1(15)}$	$\dots$		$\dots$	X_{ik15}
	g							
			1	$X_{(i+1)11}$	$\dots$	$X_{(i+1)j1}$	$\dots$	$X_{(i+1)k1}$
			$\vdots$	$\vdots$	$\ddots$			$\vdots$
		$i+1$	m			$X_{(i+1)jm}$		
			$\vdots$	$\vdots$			$\ddots$	$\vdots$
			15	$X_{(i+1)1(15)}$	$\dots$		$\dots$	$X_{(i+1)k15}$

Table E1: **Indexing Levels**. A summary of indexing levels for the WRMS estimator in combination with the logistic regression estimates (i.e., Section F) of performance measurement.

$1 \le g \le N \equiv n/2$. Players for each team are indexed by m, $1 \le m \le 15$. The covariates are indexed by j, $1 \le j \le k$. More generally, players in each game, g, $1 \le g \le N$, are indexed by m, $1 \le m \le 30$. For clarity, the player index will occasionally switch to ω, such as in the denominator of (2).

To estimate $\mathcal{W}(\boldsymbol{X})$ defined in (12), we shift the calculations away from game outcomes, i, $1 \le i \le n$, to games, g, $1 \le g \le N \equiv n/2$. This is because we assume all players in a game, g, $1 \le g \le N \equiv n/2$, are competing to amass the largest share of game value, as determined by the single-game performance measurement, Δ. By (11), we estimate Δ as the portion of win probability or fitted logit. Finally, Theorem 2.1 restricts the WRMS calculation to the set of players with playing time in a game, g, $1 \le g \le N \equiv n/2$. This set is denoted by $\mathcal{M}_g$, $1 \le g \le N \equiv n/2$, where $\#\{\mathcal{M}_g\} \le 30$. When we desire to utilize (15), there is occasion to switch the player index from a basic number index, m, $1 \le m \le 30$, to indexing by player name, π, $\pi \in \mathcal{P}$. Note that the sets $\mathcal{M}_g$ and $\overline{\mathcal{M}}_g$ may be equivalently indexed either by m, $1 \le m \le 30$, or player name, π, $\pi \in \mathcal{P}$, for any g, $1 \le g \le N$.

F Logistic Regression Additional Details

The ROI framework proposed in Figure 1 requires a performance measurement random variable or model for Δ. While many examples are possible, we propose an applied logistic regression model for performance measurement that is updated with recent player tracking data. This model is introduced briefly in Section 2.1, but the details are omitted to allow the manuscript to focus on the larger ROI framework. The present section intends to fill in these omitted details. First, the three modeling principles of aligning merit to winning, valuing as much on court activity as possible, and avoiding double counting will be detailed. Next, the initial model fitting of all 36 data fields will be presented, from which the final model of Table 1 was derived. Finally, the section will close with a robustness analysis, which finds the logistic regression model in combination with the WRMS outperforms both the Win Score and Game Score combinations with the WRMS.

F.1 Modeling Principles

We employ three principles for data selection and model calibration: aligning merit to winning, valuing as much on court activity as possible, and avoiding double counting. We now discuss each in turn.

Aligning Merit to Winning. We assume that NBA teams are attempting to maximize wins over the investment horizon. A wins-based objective function is quite standard in basketball analysis (e.g., Berri et al., 2007b, pg. 92). Other objective functions are possible, however, such as maximizing championships or maximizing operating income, see Section 4 for further discussion. Concisely, our logistic regression model is calibrated to win probability.

Valuing All Activity. From a classical statistics point-of-view, the model selection processes for exploratory observational studies often begins with data collection on a large scale (Kutner et al., 2005). As such, we desire to recognize any form of on court activity that has an effect on winning, both positive and negative. Pragmatically, this means that in addition to traditional box score categories, such as *two-point field goals made*, *turnovers*, and *blocks*, we also consider more recent player tracking and hustle statistics, such as *distance traveled*, *rebound chances*, *contested rebounds*, and *box outs*. This is an advantage of using new player tracking data in comparison to (7) and (8), though the trade-off is added complexity. In addition to data collection, we also consider this principle is selecting a logistic regression model. Specifically, we desire to recognize players with strong games despite losing at the team level. Hence, our model allows a player to make a positive individual contribution to win probability despite poor team play overall and vice versa. As a minor comment, we are at times constrained by data availability (e.g., it is preferable to track "screens set" instead of *screen assists*, but detailed data for screens set by game is not yet readily available).

Avoiding Double Counting. We desire to avoid the classic economics problem of *double counting*, which is undesirable in the measurement of macroeconomic calculations like *gross domestic product* (e.g., Mankiw, 2003, Chapter 10). In essence, our objective is to avoid giving a player double credit. For example, we create statistics such as three-point field goals missed rather than use both three-point field goals made and three-point field goal attempts. Similarly, we track two-point field goals made, three-point field goals made, and free throws made but do not also track total points scored. Other non-obvious adjustments include subtracting rebounds from *rebound chances*, subtracting blocks from *contested two-point shots*, subtracting *charges drawn* from *personal fouls drawn*, and subtracting assists, *secondary assists*, and *free throw assists* from *passes made*. In reviewing (7) and (8), we see that each equation tracks both field goals (FG) or points (PTS) and field goals attempted (FGA), which would violate this principle. Hence, the logistic regression approach we propose may offer a novel economic perspective that differs from these traditional basketball measures. In addition, these adjustments, in combination with centering each covariate, may help with issues of multicollinearity (Kutner et al., 2005).

F.2 Initial Logistic Regression Results

Our initial covariate space consists of 36 player-level statistical categories: made two-point shots (FG2O), missed two-point shots (FG2X), made three-point shots (FG3O), missed three-point shots (FG3X), made free throws (FTMO), missed free throws (FTMX), personal fouls (PF), steals (STL), adjusted offensive rebounds (i.e., offensive rebounds less contested offensive rebounds) (AORB), adjusted defensive rebounds (ADRB), assists (AST), blocks (BLKS), turnovers (TO), blocks against (BLKA), adjusted personal fouls drawn (i.e., personal fouls drawn less charges drawn) (PFD), screen assists (SAST), deflections (DEFL), charges drawn (CHGD), adjusted contested two-point shots (i.e., contested two-point shots less blocks) (AC2P), contested three-point shots (C3P), offensive box outs (OBOX), defensive box outs (DBOX), offensive loose balls recovered (OLBR), defensive loose balls recovered (DLBR), defended field goals against made (DFGO), defended field goals against missed (DFGX), drives (DRV), distance traveled in miles offense (ODIS), distance traveled in miles defense (DDIS), adjusted passes made (i.e., passes made less assists, secondary assists, and free throw assists) (APM), secondary assists (AST2), free throw assists (FAST), offensive contested rebounds (OCRB), defensive contested rebounds (DCRB), adjusted offensive rebound chances (i.e., offensive rebound chances less offensive rebounds) (AORC), and adjusted defensive rebound chances (ADRC). All adjustments are made to avoid double-counting and minimize multicollinearity concerns. For reference, a glossary of common NBA abbreviations may be found in Section C.

Model selection within statistical analysis can be a complex process (Kutner et al., 2005), often with no clear answer. We detail our approach to decide on the final model presented in Table 1 in Section 2.2. Nonetheless, in the interest of transparency and reproductive research, we also present the initial model fitting output in Table F1. Such results may provide additional insights or background, which may be used by analysts to deepen understanding of the drivers of winning in the NBA or simply explore alternative models. For reference, all data and replication code is publicly available at the repository: https://github.com/jackson-

Field	Coefficient	Standard Error	Test Statistic	Significance
(Intercept)	-0.015	0.0755	-0.20	
FG2O	0.260	0.0313	8.31	$***$
FG2X	-0.352	0.0304	-11.58	$***$
FG3O	0.551	0.0438	12.59	$***$
FG3X	-0.371	0.0297	-12.51	$***$
FTMO	0.121	0.0231	5.25	$***$
FTMX	-0.217	0.0361	-6.01	$***$
PF	-0.201	0.0231	-8.70	$***$
AORB	0.377	0.0464	8.11	$***$
ADRB	0.322	0.0259	12.44	$***$
STL	0.428	0.0401	10.67	$***$
BLK	0.128	0.0345	3.70	$***$
TOV	-0.348	0.0303	-11.49	$***$
BLKA	-0.002	0.0371	-0.04	
PFD	0.216	0.0333	6.47	$***$
AST	-0.016	0.0232	-0.68	
SAST	0.072	0.0222	3.24	$**$
DEFL	0.020	0.0202	0.99	
CHGD	0.513	0.1020	5.03	$***$
AC2P	0.041	0.0121	3.42	$***$
C3P	-0.068	0.0143	-4.77	$***$
OBOX	-0.101	0.0692	-1.46	
DBOX	0.054	0.0247	2.20	$*$
OLBR	-0.058	0.0487	-1.20	
DLBR	0.023	0.0539	0.42	
DFGO	-0.233	0.0184	-12.67	$***$
DFGX	0.076	0.0150	5.08	$***$
DRV	0.001	0.0096	0.08	
ODIS	0.094	0.2062	0.46	
DDIS	-1.104	0.2151	-5.13	$***$
APM	0.017	0.0036	4.64	$***$
AST2	0.010	0.0415	0.23	
FAST	0.010	0.0536	0.19	
OCRB	0.305	0.0387	7.87	$***$
AORC	-0.008	0.0204	-0.37	
DCRB	0.343	0.0350	9.82	$***$
ADRC	0.024	0.0151	1.59	

Table F1: **Preliminary Logistic Regression**. The initial model fitting as a first step based on team outcomes for the 2022-2023 NBA regular season. Because player tracking data was not available for four games, $n = 2{,}452$. Significant at $\alpha = 0.001$ ($***$), $\alpha = 0.01$ ($**$), and $\alpha = 0.05$ ($*$). Only fields significant at $\alpha = 0.10$ were kept in the final model of Table 1.

lautier/nba_roi.

F.3 Robustness Analysis

Recall from Section 2.1 that the underlying logistic regression model is calibrated to wins. Hence, a standard robustness analysis would be to confirm that WRMS in combination with the model of Table 1 generates output consistent with this objective. As such, we perform two types of robustness analysis.

The first is to compare the actual team wins of the 2022-2023 NBA regular season against the team total of (12), (13), and (14). In other words, because

$$\sum_{g=1}^{N} \sum_{m \in \mathcal{M}_g} \mathcal{W}(\mathcal{S})_{gm} = N,$$

by definition, it is desirable to compare how many wins are allocated to each team by each model with the actual number of wins recorded by each team for the 2022-2023 NBA regular season. We do exactly this in Table F2. Recall $n = 2{,}452$, which implies there are 1,226 wins to be allocated (four games from the 2022-2023 NBA regular season were missing player tracking data). The reported average absolute errors are larger than the now dated 1.67 observed in Berri et al. (2007b, Table 6.8). The standardization tends to pull teams towards the center, and so the larger errors are generally at the very top and bottom of the standings. Of (12), (13), and (14), the logistic regression is the most accurate for both average and median absolute errors by either win total or team rank. One interpretation of these results is that the logistic regression, thanks to its initial calibration to wins, is more attuned to winning than either Game Score or Win Score. On the other hand, the results are comparable, which is impressive given the simplicity of the Game Score and Win Score formulas. Of course, with modern data collection methods and statistical software, the effort necessary to generate the logistic regression estimates is minimal (recall also that all data and replication code is publicly available at the repository: https://github.com/jackson-lautier/nba_roi).

As a second validation, we perform a logistic regression against game outcome using a team's single game total of (12), (13), and (14). We find that both a team's total $\mathcal{W}(\boldsymbol{X})$ and WnSc* are highly significant to increase team win probability. GmSc* is not significant, though it is likely due to WnSc* and GmSc* being highly correlated. The most significant is $\mathcal{W}(\boldsymbol{X})$ based on a standard variable importance analysis (Kuhn, 2008). This is likely due to the fact that $\mathcal{W}(\boldsymbol{X})$ uses many more data fields than either GmSc* or WnSc*. In any subset combination of two, both models each register coefficients as highly significant. In a standard variable importance analysis (Kuhn, 2008), $\mathcal{W}(\boldsymbol{X})$ always registers as the most important. In a model using only GmSc* and WnSc*, WnSc* registers as the most important. The results of Tables F2 and F3 simultaneously indicate that all three models (12), (13), and (14) have merits, of which $\mathcal{W}(\boldsymbol{X})$ has the strongest connection to winning (followed by WnSc* and then GmSc*).

G Performance Measurement Comparisons

The motivation for the flexibility of (5) is a *plug and play* attribute of the proposed ROI framework. For example, it is possible to select any performance measurement of on court basketball performance that is calibrated to a single game for Δ. As we illustrate with Figure 2, this choice can have a significant influence on the dollar allocation of SGV to each player. The purpose of the present section is to provide additional detail on the comparison of player performance for (12), (13), and (14) as it relates to (16).

Figure G1 presents an aggregated comparison of (12), (13), and (14) as it relates to (16) by comparing player percentiles. The off-diagonals show significant disagreements in player performance, especially between PVWL and either PVWS and PVGS. One explanation for these differences is that the model of Table 1 uses player tracking data, which allows for more detail than either (7) or (8). For example, the model of Table 1 does not report assists (AST) as significant but instead finds adjusted passes made (APM) as significant. In comparing PVWS and PVGS, we see general similarities. This may suggest limited differences in these

Rank	Team	Wins	WL (ae)	WS (ae)	GS (ae)	WLR (ae)	WSR (ae)	GSR (ae)
	Median Error		3.66	4.95	4.82	1.00	3.00	4.00
	Average Error		5.49	5.99	6.47	2.87	3.93	4.87
1	MIL	58	46.08 (11.9)	45.08 (12.9)	42.13 (15.9)	1 (0)	2 (1)	9 (8)
2	BOS	57	45.78 (11.2)	45.60 (11.4)	43.71 (13.3)	2 (0)	1 (1)	2 (0)
3	PHI	54	45.22 (8.8)	42.81 (11.2)	42.40 (11.6)	5 (2)	7 (4)	6 (3)
4	DEN	53	45.61 (7.4)	44.71 (8.3)	43.52 (9.5)	3 (1)	3 (1)	3 (1)
5	MEM	51	44.44 (6.6)	43.69 (7.3)	42.95 (8.0)	6 (1)	5 (0)	5 (0)
6	CLE	51	42.03 (9.0)	40.89 (10.1)	41.03 (10.0)	10 (4)	18 (12)	18 (12)
7	SAC	48	45.60 (2.4)	44.57 (3.4)	43.89 (4.1)	4 (3)	4 (3)	1 (6)
8	NYK	47	41.19 (5.8)	41.77 (5.2)	41.42 (5.6)	18 (10)	11 (3)	12 (4)
9	BKN	45	42.46 (2.5)	41.31 (3.7)	41.15 (3.8)	9 (0)	13 (4)	16 (7)
10	PHX	45	42.90 (2.1)	41.13 (3.9)	41.12 (3.9)	7 (3)	15 (5)	17 (7)
11	LAC	44	42.03 (2.0)	40.89 (3.1)	40.27 (3.7)	11 (0)	17 (6)	22 (11)
12	MIA	44	36.64 (7.4)	37.89 (6.1)	38.95 (5.1)	27 (15)	26 (14)	25 (13)
13	GSW	43	41.62 (1.4)	42.86 (0.1)	42.29 (0.7)	14 (1)	6 (7)	7 (6)
14	LAL	43	41.96 (1.0)	42.74 (0.3)	42.22 (0.8)	12 (2)	8 (6)	8 (6)
15	NOP	42	41.56 (0.4)	41.27 (0.7)	41.40 (0.6)	15 (0)	14 (1)	14 (1)
16	ATL	41	41.24 (0.2)	42.69 (1.7)	43.10 (2.1)	17 (1)	9 (7)	4 (12)
17	MIN	41	40.26 (0.7)	40.00 (1.0)	40.54 (0.5)	21 (4)	22 (5)	20 (3)
18	TOR	41	39.23 (1.8)	40.02 (1.0)	41.42 (0.4)	22 (4)	21 (3)	13 (5)
19	OKC	40	40.99 (1.0)	40.75 (0.8)	41.59 (1.6)	19 (0)	19 (0)	11 (8)
20	CHI	39	40.51 (1.5)	41.00 (2.0)	40.52 (1.5)	20 (0)	16 (4)	21 (1)
21	DAL	38	41.36 (3.4)	39.01 (1.0)	39.38 (1.4)	16 (5)	23 (2)	23 (2)
22	UTA	37	41.79 (4.8)	41.68 (4.7)	41.33 (4.3)	13 (9)	12 (10)	15 (7)
23	WAS	35	42.87 (7.9)	41.82 (6.8)	40.92 (5.9)	8 (15)	10 (13)	19 (4)
24	IND	35	38.34 (3.3)	40.28 (5.3)	41.67 (6.7)	24 (0)	20 (4)	10 (14)
25	ORL	34	37.31 (3.3)	38.22 (4.2)	38.60 (4.6)	25 (0)	24 (1)	27 (2)
26	POR	33	36.96 (4.0)	38.21 (5.2)	39.24 (6.2)	26 (0)	25 (1)	24 (2)
27	CHA	27	35.09 (8.1)	37.87 (10.9)	38.83 (11.8)	28 (1)	27 (0)	26 (1)
28	HOU	22	38.59 (16.6)	36.92 (14.9)	37.20 (15.2)	23 (5)	28 (0)	28 (0)
29	SAS	21	33.67 (12.7)	35.96 (15.0)	37.05 (16.1)	29 (0)	29 (0)	29 (0)
30	DET	17	32.68 (15.7)	34.37 (17.4)	36.18 (19.2)	30 (0)	30 (0)	30 (0)

Table F2: **Model Versus Actual Wins**. A comparison of actual versus estimated wins using the $\mathcal{W}(\boldsymbol{X})$ (WL) (12), the Game Score (GS) (13), and the Win Score (WS) (14) models. The absolute errors (ae) are included, and we also report the model rankings (WLR, WSR, GSR) versus the actual team ranking. All results are for the 2022-2023 NBA regular season. The actual wins are adjusted to omit games without player tracking data available (GSW, CHI, MIN, and SAS).

two approaches. For a summary of the top disagreements between sum totals of (12), (13), and (14) along the lines of (16), see Table G1. For complete results, navigate to the public `github` repository at https://github.com/jackson-lautier/nba_roi.

Field	Coefficient	Standard Error	Test Statistic	Significance
(Intercept)	-14.278	0.6328	-22.56	$***$
$\mathcal{W}(\boldsymbol{X})$	17.811	1.1961	14.89	$***$
WnSc*	10.502	2.5387	4.14	$***$
GmSc*	0.884	2.2568	0.39	

Table F3: **Team Level Models and Wins**. A logistic regression using team totals of (12), (13), and (14) against the game outcome for the total sample of 2,452 game outcomes for the 2022-2023 NBA regular season. Significant at $\alpha = 0.001$ ($***$), $\alpha = 0.01$ ($**$), $\alpha = 0.05$ ($*$), and $\alpha = 0.10$ ($\cdot$). The McFadden R^2 (McFadden, 1974) is 0.5203. WnSc* and GmSc* are highly correlated, and any subset logistic regression with any combination of two reports each model coefficient as significant at $\alpha = 0.001$ ($***$).

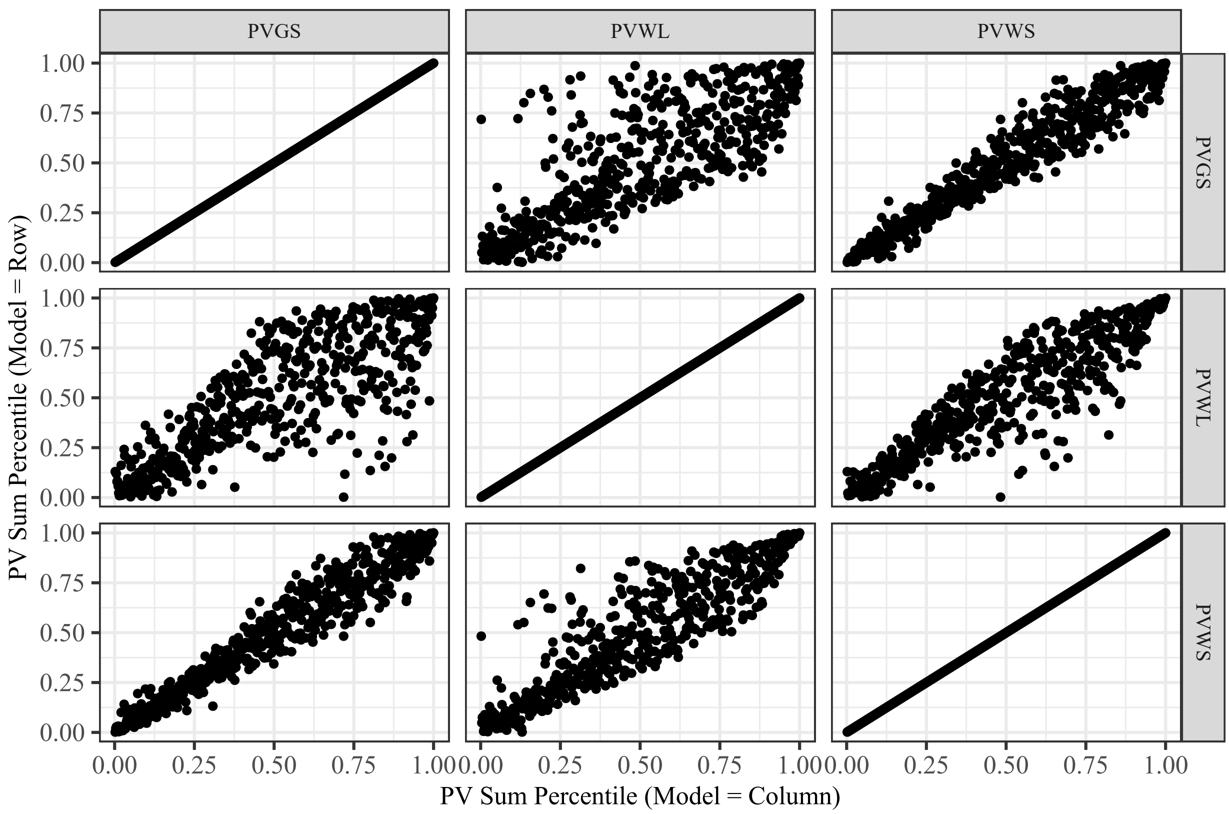

Figure G1: **PVW(·) Percentile Comparisons**. Percentile plots between sum totals of (12) (WL), (13) (GS), and (14) (WS) for the 2022-2023 NBA regular season (i.e., (16)) in terms of percentile rank (%). The further a plot deviates from a straight line, the more disagreement between players.

H Simulation Study

We first conduct a simulation study to verify consistency of the WRMS, (i.e., property (*ii*) of Theorem 2.1). We assume a sample of $N = 1{,}000$ games, with each team playing between

Name	WL(%)	WS(%)	Name	WL(%)	GS(%)	Name	WS(%)	GS(%)
CJ McCollum	0.31	0.82	Dillon Brooks	0.00	0.72	Jordan Poole	0.66	0.91
Anfernee Simons	0.16	0.65	Anfernee Simons	0.16	0.85	Jaden Ivey	0.55	0.80
Terry Rozier	0.20	0.69	Terry Rozier	0.20	0.87	Jalen Green	0.68	0.92
Dillon Brooks	0.00	0.48	Jaden Ivey	0.14	0.80	Dillon Brooks	0.48	0.72
Killian Hayes	0.12	0.54	Jalen Green	0.28	0.92	Isaiah Hartenstein	0.87	0.65
Jaden Ivey	0.14	0.55	CJ McCollum	0.31	0.94	Andre Drummond	0.79	0.57
Jordan Clarkson	0.21	0.62	Jordan Clarkson	0.21	0.83	Jordan Clarkson	0.62	0.83
Jalen Green	0.28	0.68	Killian Hayes	0.12	0.72	Steven Adams	0.83	0.63
LaMelo Ball	0.22	0.62	RJ Barrett	0.28	0.84	Usman Garuba	0.65	0.45
Fred VanVleet	0.47	0.86	LaMelo Ball	0.22	0.76	Anfernee Simons	0.65	0.85

Table G1: **Player Performance Disagreements**. The top ten largest disagreements between sum totals of (12) (WL), (13) (GS), and (14) (WS) for the 2022-2023 NBA regular season (i.e., (16)) in terms of percentile rank (%).

1 and 5 players (10 total). The number of players appearing for each team is a discrete uniform random variable over the integers $\{1, \ldots, 5\}$. Furthermore, the performance random variable for each player follows an i.i.d. exponential distribution with rate parameter equal to 1, denoted exp(1). The simulation procedure is

1. Simulate 1,000 $\times$ 10 i.i.d. exp(1) random variables.

2. For each game, $g = 1, \ldots, 1{,}000$, simulate two discrete uniform random variables over $\{1, \ldots, 5\}$ to determine how many players appear for each team.

3. For each game, $g = 1, \ldots, 1{,}000$, calculate the natural share, as defined by (2), using the simulated i.i.d. exp(1) random variables from Step 1.

4. For each player, $m \in \mathcal{M}_g$, appearing in each game, g, $1 \leq g \leq 1{,}000$, we calculate $\mathcal{W}$.

5. For each player, $m \in \mathcal{M}_g$, appearing in each game, g, $1 \leq g \leq 1{,}000$, we calculate the bias by subtracting the calculated natural share in Step 3 from the calculated $\mathcal{W}$ in Step 4.

From our sample, we obtain $m^* = 6{,}081$, $\bar{m} = 6.081$, $\bar{\Delta}_{m^*} = 0.9939$, and $s(\Delta)_{m^*} = 0.9861$. This results in an empirical mean bias of 0.0000 over the simulated sample of 6,081 players (the empirical median bias is 0.0007). This is numerical verification of Theorem 2.1, (*ii*).

We next provide a simulation study to verify the results of Theorem 3.1. We estimate (15) using (12) for all $g = 1, \ldots, n/2$ and $\pi \in \mathcal{P}$ using data from the 2022-2023 NBA regular season. These estimates correspond to Section 2.2. Thus, $n = 2{,}452$. Further, we assume $\text{SGV}_g \sim \mathcal{N}(\mu = 100, \sigma^2 = 25)$ for all $g = 1, \ldots, 1{,}226$. We run the following simulation for 1,000 replicates. That is, for each replicate, $r = 1, \ldots, 1{,}000$:

1. Simulate 1,226 random variables from a $\mathcal{N}(\mu = 100, \sigma^2 = 25)$ distribution, which we denote by $\widehat{\text{SGV}}_g$, $g = 1, \ldots, 1{,}226$.

2. Compute the product

$$\hat{\theta}_g = \widehat{\text{SGV}}_g \sum_{\pi \in \overline{\mathcal{M}}_g} \mathcal{W}(\boldsymbol{X})^*_{g\pi},$$

   for $g = 1, \ldots, 1{,}226$.

3. Save the result as the summation,

$$\text{Result}_r = \sum_{g=1}^{1{,}226} \hat{\theta}_g.$$

In doing so, we find an empirical mean of

$$\frac{1}{1{,}000} \sum_{r=1}^{1{,}000} \text{Result}_r = 122{,}605.6,$$

which is quite close to $\mu(n/2) \equiv 100 \times 1{,}226$. In Figure H1, we provide a density plot of the simulated results.

Next, we state a minor extension to Theorem 3.1.

Result C.1. *Assume the conditions of Theorem 3.1, and further assume* $\text{Var}(\text{SGV}_g) = \sigma^2$ *for all* $g = 1, \ldots, N \equiv n/2$. *If* SGV_g *is independent of* SGV_{g^*} *for all* $g, g^* = 1, \ldots, n/2$, $g \neq g^*$, *then*

$$\text{Var}\left(\sum_{g=1}^{n/2} \sum_{\pi \in \overline{\mathcal{M}}_g} \text{SGV}_g \mathcal{W}^*_{g\pi} \middle| \mathcal{W}^*_{g\pi} \right) = \sigma^2 \sum_{g=1}^{n/2} \left(\sum_{\pi \in \overline{\mathcal{M}}_g} \mathcal{W}^*_{g\pi} \right)^2.$$

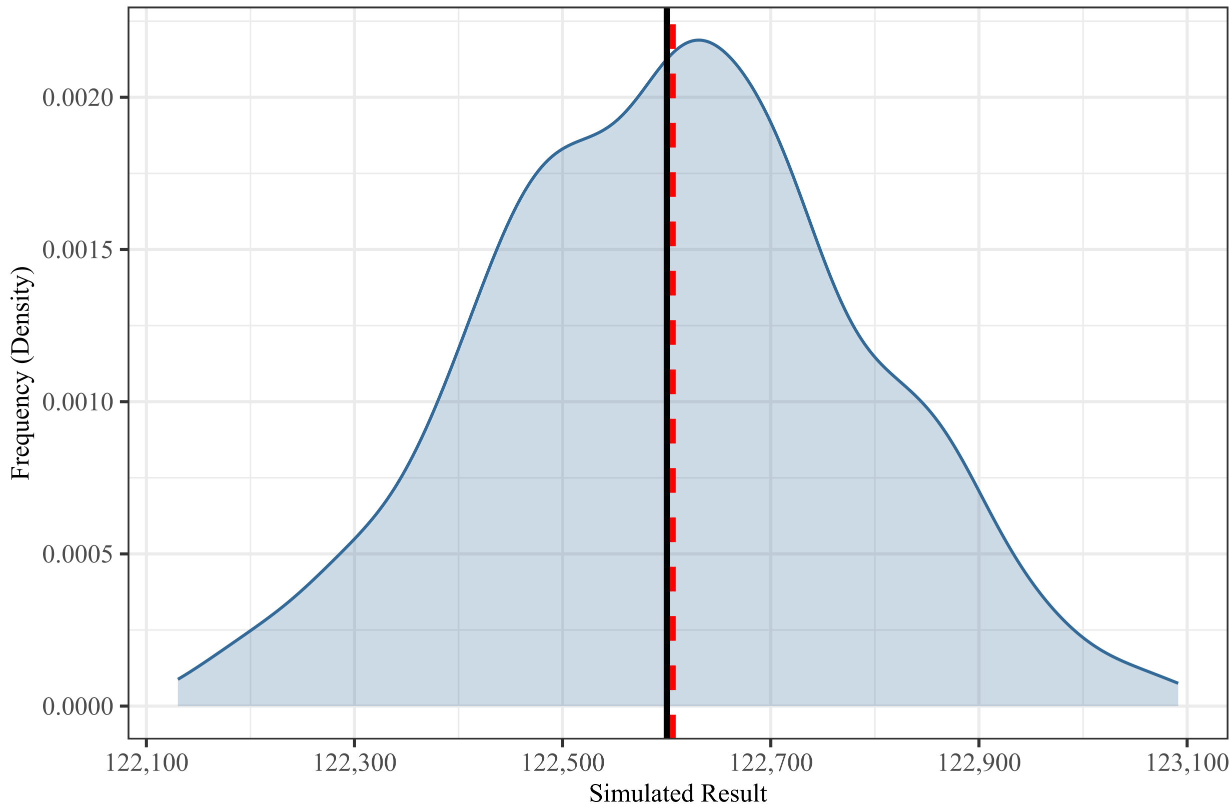

Figure H1: **Simulation Study Results**. A density plot of 1,000 replicates to verify Theorem 3.1. The vertical black line indicates the theoretical mean using Theorem 3.1. The vertical dashed line indicates the empirical sample mean of the 1,000 replicates. The two quantities are quite close, which is a simulation validation of Theorem 3.1.

Proof. By independence,

$$\begin{aligned}
\mathrm{Var}\left(\sum_{g=1}^{n/2}\sum_{\pi\in\overline{\mathcal{M}}_g}\mathrm{SGV}_g\mathcal{W}^*_{g\pi}\middle|\mathcal{W}^*_{g\pi}\right) &= \sum_{g=1}^{n/2}\mathrm{Var}\left(\mathrm{SGV}_g\sum_{\pi\in\overline{\mathcal{M}}_g}\mathcal{W}^*_{g\pi}\middle|\mathcal{W}^*_{g\pi}\right)\\
&= \sum_{g=1}^{n/2}\left(\sum_{\pi\in\overline{\mathcal{M}}_g}\mathcal{W}^*_{g\pi}\right)^2\mathrm{Var}(\mathrm{SGV}_g)\\
&= \sigma^2\sum_{g=1}^{n/2}\left(\sum_{\pi\in\overline{\mathcal{M}}_g}\mathcal{W}^*_{g\pi}\right)^2.
\end{aligned}$$

□

In an additional simulation study with 10,000 replicates, we obtain an empirical sample variance of the results vector, $\{\text{Result}_r\}_{1 \leq r \leq 10{,}000}$, of 32,414.45. This is quite close to the true result, which we calculate to be 31,119.83.

Finally, we verify the results of Section D with a simulation study. In this instance, we assume a sample of $N = 1{,}000$ games, with each team playing a nonrandom 5 players. The number of players is held fixed to verify the results of Section D. Further, we assume the i.i.d. performance random variables are $\Delta = -0.25\rho_1 + 0.25\rho_2$, where $\rho_1 \sim \mathcal{N}(\mu = 0, \sigma = 5)$ and $\rho_2 \sim \mathcal{N}(\mu = 0, \sigma = 7)$. Thus, the natural share defined in (2) follows (S.1) with $\sigma_x^2 = 5^2/16 + 7^2/16 = 4.625$ and $\sigma_y^2 = 9\sigma_x^2$. To verify this with simulation, we

1. Simulate $1{,}000 \times 10$ i.i.d. $\Delta = -0.25\rho_1 + 0.25\rho_2$ random variables.

2. For each game, $g = 1, \ldots, 1{,}000$, calculate the natural share, as defined by (2), using the simulated i.i.d. exp(1) random variables from Step 1.

3. Simulate 10,000 Cauchy random variables with location parameter $x_0 = 0.10$ and scale parameter $\gamma = 0.3$ per (S.1).

4. Compare a QQ-plot of the middle 90% of the ordered 10,000 observations from Step 2 and the ordered 10,000 observations from Step 3. We use the middle 90% because of the tendency for extreme observations from the Cauchy distribution. The results may be found in Figure H2, which indicates numerical validation of the result of Section D.

References

J. Berk and P. Demarzo (2007). *Corporate Finance, 1st Edition*. Pearson Education, Inc.

D. J. Berri (1999). "Who is 'most valuable'? Measuring the player's production of wins in the National Basketball Association." *Managerial and Decision Economics* **20**, 411–427.

D. J. Berri and J. C. Bradbury (2010). "Working in the land of the metricians." *Journal of Sports Economics* **11**, 29–47.

D. J. Berri, S. L. Brook, B. Frick, A. J. Fenn and R. Vicente-Mayoral (2005). "The short supply of tall people: Competitive imbalance and the National Basketball Association." *Journal of Economic Issues* **39**, 1029–1041.

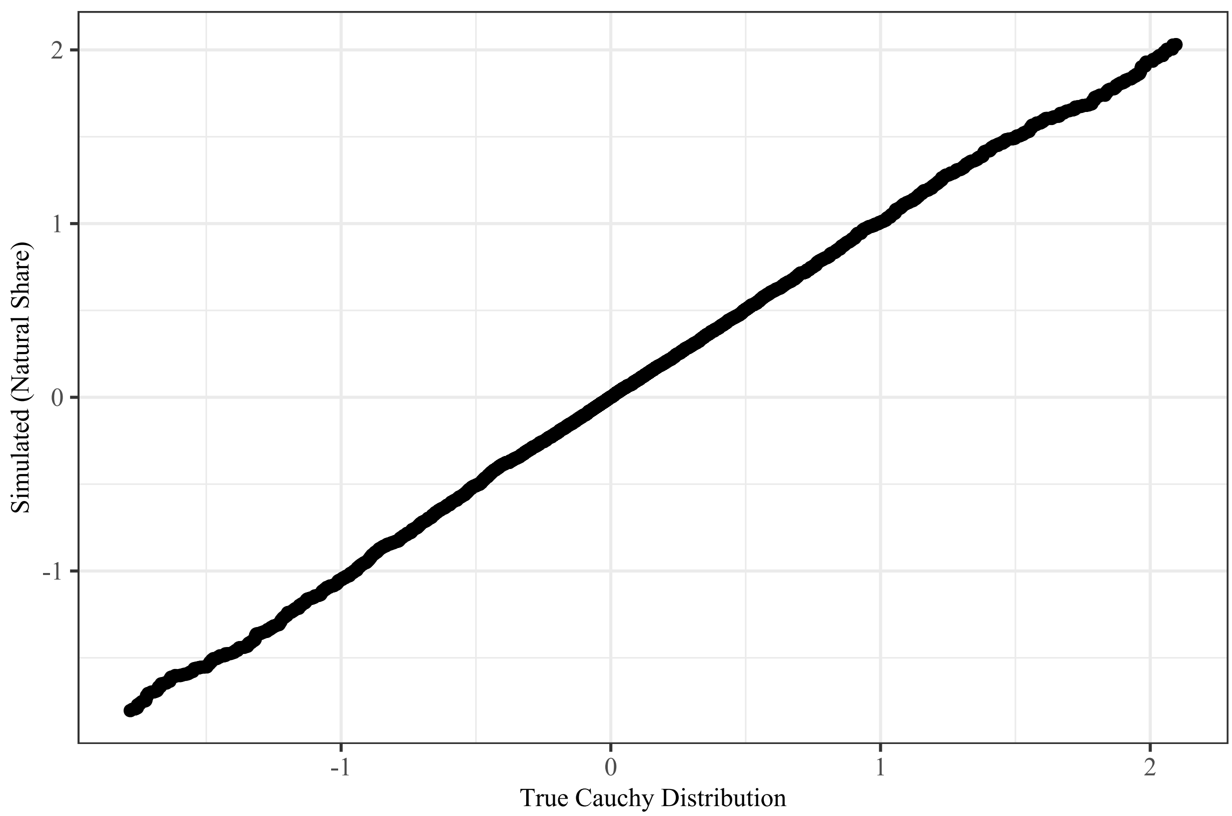

Figure H2: **Cauchy Simulation Results**. A QQ-plot of the middle 90% of ordered data from simulated natural shares in the form of a ratio of independent normal random variables and a Cauchy distribution with location and scale parameters per (S.1). The closeness of the distributions represents simulation verification of the result of Section D.

D. J. Berri, S. L. Brook and M. B. Schmidt (2007a). "Does one simply need to score to score?" *International Journal of Sport Finance* **2**, 190–205.

D. J. Berri and A. C. Krautmann (2006). "Shirking on the court: Testing for the incentive effects of guaranteed pay." *Economic Inquiry* **44**, 536–546.

D. J. Berri, M. B. Short and S. L. Brook (2007b). *The Wages of Wins: Taking Measure of the Many Myths in Modern Sport*. Standford Business Books.

M. Casals and J. A. Martínez (2013). "Modelling player performance in basketball through mixed models." *International Journal of Performance Analysis in Sport* **13**, 64–82.

P. Fearnhead and B. M. Taylor (2011). "On estimating the ability of NBA players." *Journal of Quantitative Analysis in Sports* **7**.

N. Halevy, E. Y. Chou, A. D. Galinsky and J. K. Murnighan (2012). "When hierarchy wins: Evidence from the National Basketball Association." *Social Psychological and Personality Science* **3**, 398–406.

T. Idson and L. Kahane (2000). "Team effects on compensation: An application to salary determination in the National Hockey League." *Economic Inquiry* **38**, 345–357.

J. Kuehn (2017). "Accounting for complementary skill sets: Evaluating individual marginal value to a team in the National Basketball Association." *Economic Inquiry* **55**, 1556–1578.

M. Kuhn (2008). "Building predictive models in R using the caret package." *Journal of Statistical Software* **28**, 1–26.

M. H. Kutner, C. J. Nachtsheim, J. Neter and W. Li (2005). *Applied Linear Statistical Models*. McGraw-Hill Irwin.

N. G. Mankiw (2003). *Principles of Macroeconomics, 3rd Edition*. South-Western College Pub.

J. A. Martínez (2012). "Factors determining production (FDP) in basketball." *Economics and Business Letters* **1**, 21–29.

D. McFadden (1974). "Conditional logit analysis of qualitative choice behavior." *Frontiers in Econometrics* pp. 105–142.

N. Mukhopadhyay (2000). *Probability and Statistical Inference*. New York, NY: Marcel Dekker.

National Basketball Association (2018). "Highlights of the Collective Bargaining Agreement between the National Basketball Association (NBA) and National Basketball Players Association (NBPA)." nba.com. `https://cdn.nba.com/manage/2021/03/2018-19-CBA.pdf`. Online; accessed 19 July 2023.

National Basketball Association (2023). "Statistics." nba.com. `https://www.nba.com/stats`. Online; accessed 13 June 2023.

G. L. Page, G. W. Fellingham and C. S. Reese (2007). "Using box-scores to determine a position's contribution to winning basketball games." *Journal of Quantitative Analysis in Sports* **3**.

R. Simmons and D. J. Berri (2011). "Mixing the princes and the paupers: Pay and performance in the National Basketball Association." *Labour Economics* **18**, 381–388. The Economics of Sports Labour Markets.

R. Tunaru, E. Clark and H. Viney (2005). "An option pricing framework for valuation of football players." *Review of Financial Economics* **14**, 281–295. Real Options SI.