sociological science

Supplement to:

McMahan, Peter, and Eran Shor. 2024. "Status ambiguity and multiplicity in the selection of NBA awards." Sociological Science 11: 680-706.

Appendix A: Data errors and miscalculations

In replicating Biegert, Kühhirt, and Van Lancker's (2023) analysis using the data and code they provided, we encountered substantial problems with the data and calculations used in their analysis (available at https://osf.io/ntwdy/). We have identified four major types of errors in the authors' data: (1) temporal misalignment of variables resulting in the assignment of values to the wrong season, (2) non-existent (phantom) player–seasons that are included in the data, (3) players' seasons that are missing from the data, and (4) miscalculated averages and aggregations.

The first substantial issue lies in the temporal misalignment of variables in the dataset. This issue stems from Biegert, Kühhirt, and Van Lancker 's non-standard and inappropriate method of creating time-lagged variables. As a result, most of the data aligns with one season earlier than it should—for example data for the 2002–2003 season was labeled as 2001–2002, and so on. The effects of this misalignment on the final analysis are reduced somewhat because the misalignment is applied consistently to nearly all the variables used in the models, resulting in a parallel realignment of the data. However, variables relating to playoff appearances remain coded in their correct season and are therefore misaligned with the rest of the data. In addition, this misalignment contributed to two additional issues: the phantom player–seasons and the missing player–seasons, which we discuss in the next two paragraphs.

The second, and arguably most problematic, issue we encountered was the inclusion in the dataset of phantom player–seasons, i.e., player–seasons that are used in the analysis but represent seasons in which a player was in fact not playing in the NBA and not eligible for All-Star selection. Of the 10,188 player–seasons included in Biegert, Kühhirt, and Van Lancker 's data, 1,265 (12.42%) represent seasons in which the player was not in fact in the league. While most of these phantom seasons are appended to the end of a player's career, others represent a season in which the player was injured, retired, or already played in a different league. Each of these phantom player–seasons is treated by Biegert, Kühhirt, and Van Lancker as a year in which the player was not selected for the All-Star game, while in fact the players were not eligible for All-Star selection in these years, as they were out of the league.

A related issue with the data is the failure to include some existing player–seasons—seasons in which players were in the league but are nevertheless absent from the dataset. Of the 11,746 player–seasons between 1984 and 2016 available on Basketball Reference, 9,361 have full lagged data necessary to analyze cumulative status effects (note that, similarly to Biegert, Kühhirt, and Van Lancker, this includes dropping every player's rookie season, as they were not eligible for All-Star game selection before the start of their NBA career). Of these, 438 (4.68%) are missing from Biegert, Kühhirt, and Van Lancker's data. Even in their final regression, measuring cumulative status bias (model 6), which includes many more variables and therefore more missing values, the analysis excludes 143 (1.53%) player–seasons that had full data available.

The last of the coding issues, affecting variables throughout the dataset, concerns the averaging of players' statistics across games. Standard sources like Basketball Reference calculate statistics such as points per 36 minutes across a defined time period by tallying the total number of points over all relevant games and normalizing it using the total number of minutes played across those

games. In contrast, Biegert, Kühhirt, and Van Lancker calculate the points per 36 minutes independently for each relevant game, averaging across all of these games. While the difference between these methods is sometimes negligible, it often leads to stark variations in players' season averages (see the bottom two panels of figure 1A below). This is especially notable for players who played only a few minutes per game. For example, a bench player could score four points in the last two minutes of a decided game and yield a game-level rate of 72 points per 36 minutes. Such a fantastically high score in one game can strongly skew the mean calculated across games in an entire season.

To illustrate the issues described above, Figure 1A compares Biegert, Kühhirt, and Van Lancker's points per 36 minutes statistic for four NBA players with the data calculated from BR. In the figure, we re-align the seasons in Biegert, Kühhirt, and Van Lancker's data to correspond with the corrected data for visual clarity, but it is important to keep in mind that BKLV's data is in fact shifted one year to the left. The top two panels represent very well-known and highly valued players, Michael Jordan and Tim Hardaway, both selected multiple times to the NBA All-Star game. The bottom two panels trace the careers of two less successful players, Earl Barron and Doug Lee. Phantom seasons in Biegert, Kühhirt, and Van Lancker's data are demonstrated in all four panels, including, for example, Jordan's 1993-94 season, when he was out of the league playing minor league baseball and ineligible for All-Star selection, and an exceptionally high score for the 1993-94 season for Hardaway, when he was in fact not playing due to a knee injury. Notably, all four players also include a post-career phantom season in Biegert, Kühhirt, and Van Lancker's data. The statistics from Biegert, Kühhirt, and Van Lancker are also missing for several seasons, including Jordan's 1995-96 and 2001-02 seasons (he was selected to the All-Star game in both) and Barron's 2011 season.

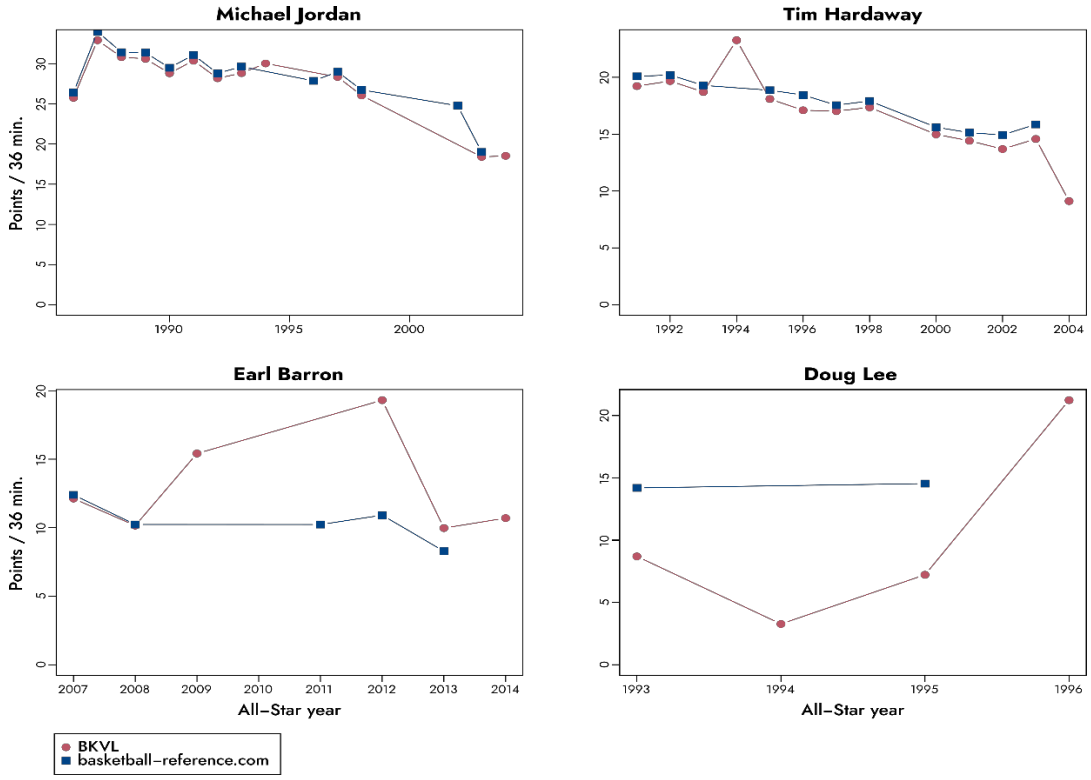

Figure 1A: Points per 36 minutes across the career of four NBA players. Red circles show the mean point calculations by Biegert, Kühhirt, and Van Lancker (2023). Blue squares show the mean points as calculated by the authors of the present study directly from baskeball-reference.com. Horizontal axis values represent the year of the All-Star election. The figure only includes player–seasons included in the regression analyses, meaning that seasons without lagged data (such as each players' rookie season) are not included.

For the seasons that are correctly included, Figure 1A also illustrates the divergence between Biegert, Kühhirt, and Van Lancker and our own calculated averages. The difference is minimal for Jordan and Hardaway, who averaged at least 30 minutes per game for most of their career. However, for more marginal players, such as Barron and Lee, sporadic and usually short stints on the court lead to higher game-to-game variability in points per 36 minutes. Consequently, we see a greater disparity between our own calculations and those of Biegert, Kühhirt, and Van Lancker, with the latter showing high variability and uncommonly high-scoring seasons, falsely suggesting that an All-Star selection may have been merited.