Challenging nostalgia and performance metrics in baseball

Daniel J. Eck
Department of Biostatistics, Yale University, New Haven, CT, 06510.
daniel.eck@yale.edu

1 Introduction

It is easy to be blown away by the accomplishments of great old time baseball players when you look at their raw or advanced baseball statistics. These players produced mind-boggling numbers. For example, see Babe Ruth's batting average and pitching numbers, Honus Wagner's 1900 season, Ty Cobb's 1911 season, Walter Johnson's 1913 season, Tris Speaker's 1916 season, Rogers Hornsby's 1925 season, and Lou Gehrig's 1931 season. The statistical feats achieved by these players (and others) far surpass the statistics that recent and current players produce. At first glance it seems that players from the old eras were vastly superior to the players in more modern eras, but is this true? Were the old timers actually better? In this paper, we investigate whether baseball players from earlier eras of professional baseball are overrepresented among the game's all-time greatest players according to popular opinion, performance metrics, and expert opinion. We define baseball players from earlier eras to be those that started their MLB careers in the 1950 season or before. We chose this year because it coincides with the decennial US Census and is close to 1947, the year in which baseball became integrated.

In this paper we do not compare baseball players via their statistical accomplishments. Such measures exhibit era biases that are confounded with actual performance. Consider the single season homerun record as an example. Before Babe Ruth, the single season homerun record was 27 by Ned Williams in 1884 Babe Ruth broke this record in 1919 when he hit 29 homeruns. He subsequently destroyed his own record in the following 1920 season when he hit 54 homeruns. The runner up in 1920 finished the season with a grand total of 15 homeruns. At this point in time homerun hitting was not an integral part of a batter's approach. This has changed. Now, we often see multiple batters reach at least 30-40 homeruns within one season and a 50 homerun season is not a rare occurrence. In the 1920s, Babe Ruth stood head and shoulders above his peers due to a combination of his innate talent and circumstance. His approach was quickly emulated and became widely adopted. However, Ruth's accomplishments as a homerun hitter would not stand out nearly as much if he played today and put up similar homerun totals. The example of homeruns hit by Babe

Ruth and the impact they had relative to his peers represents a case where adjustment towards a peer-derived baseline fails across eras. No one reasonably expects 1920 Babe Ruth to hit more than three times the amount of homeruns hit by the second best homerun hitter if the 1920 version of Babe Ruth played today. This is far from an isolated case.

There are several statistical approaches used to compare baseball players across eras. Examples include wins above replacement as calculated by baseball reference (bWAR), wins above replacement as calculated by fangraphs (fWAR), adjusted OPS+, adjusted ERA+, era-adjusted detrending (Petersen et al., 2011), computing normal scores as in Jim Albert's work on a Baseball Statistics Course in the Journal of Statistics Education, and era bridging (Berry et al., 1999). A number of these are touted to be season adjusted and the remainder are widely understood to have the same effect. In one way or another all of these statistical approaches compare the accomplishments of players within one season to a baseline that is computed from statistical data within that same season. This method of player comparison ignores talent discrepancies that exist across seasons as noted by Stephen J. Gould in numerous lectures and papers. Currently, there is no definitive quantitative or qualitative basis for comparing these baselines, which are used to form intra-season player comparisons, across seasons. These methods therefore fail to properly compare players across eras of baseball despite the claim that they are season adjusted.

Worse still is that these approaches exhibit a favorable bias towards baseball players who played in earlier seasons (Schmidt and Berri, 2005). We explore this bias from two separate theoretical perspectives underlying how baseball players from different eras would actually compete against each other. The first perspective is that players would teleport across eras to compete against each other. From this perspective, the players from earlier eras are at a competitive disadvantage because, on average, baseball players have gotten better as time has progressed. Specifically, it is widely acknowledged that fastball velocity, pitch repertoire, training methods, and management strategies have all improved over time. We do not find the teleportation perspective to be of much interest for these reasons. The second perspective is that a player from one era could adapt naturally to the game conditions of another era if they grew up in that time. This line of thinking is challenging to current statistical methodology because adjustment to a peer-derived baseline no longer makes sense.

Even in light of these challenges with the second perspective, we find that the players from earlier eras are overrepresented among baseball's all time greats. We justify our findings through the consideration of population dynamics which have changed drastically over time.

2 Data

The eligible MLB population is not well-defined. As a proxy, we can say that the eligible MLB population is the decennial count of males aged 20-29 that are living in the United States (collected on years 1880, 1890,...) and Canada (collected on years 1881, 1891,...). This information is readily obtainable and does not explicitly double count individuals over the course of its collection. The MLB eligible population is displayed in Table 1. The cumulative proportion means that at each era, the population of the previous eras is also included. As an example of how to interpret this dataset, consider the year 1950. There were 11.59 million males aged 20-29. The proportion of the historical eligible MLB population that existed at or before 1950 is 0.187.

We now explain the specifics of MLB eligible population data recorded in Table 1. The MLB started in 1876 so our data collection begins with the 1880 Census. Baseball was finally integrated in 1947. As a result, African American and Hispanic demographic data is added to our dataset starting in 1960. The year 1960 is chosen because the integration of the MLB was slow as noted in Armour's work on the integration of baseball in the Society for American Baseball Research. We obtain this demographic data for the United States eligible MLB population from the US Census for years 1880-1950 and from Statistics Canada for years 1881-1951. From 1960-2010, we use the United Nations Census for the United States eligible MLB population. The latter census is also used to estimate the population in the global talent pool from which the MLB draws. We note that the black population in Canada is relatively negligible for time periods before 1960 as noted in Milan and Tran's work in Canadian Social Trends Spring. Therefore these tallies are not included in the eligible MLB population before 1960. The same is done for the Latin American population in Canada.

African American and Hispanic US citizens were not the only groups discriminated against by pre- (and post-) integration baseball. Players from Central and South American countries and the Caribbean islands were also discriminated against. Citizens from those countries are added to the eligible MLB player pool in 1960. The population data for these countries are obtained from the UN. The countries included in our dataset are Mexico, the Dominican Republic, Venezuela, Cuba, Panama, Puerto Rico, Netherlands Antilles, Aruba, Honduras, Jamaica, the Bahamas, Peru, Columbia, Nicaragua, and the United States Virgin Islands. In the 2000s, the MLB and minors saw an influx of Asian baseball players from Japan, South Korea, Taiwan, and the Philippines. The populations from these countries are added to the eligible MLB player talent pool for those years. We obtain the Japanese, South Korean, and Philippines census data from the UN The census data from Taiwan is estimated from the CIA World Factbook, the same age stratification was not obtainable so we use the population of males ages 0-14 for the 2010 Taiwanese MLB eligible population and males ages 15-24 for the 2000 Taiwanese MLB eligible population. In 2010, the MLB established a national training center in Brazil as noted in Loré's work on the popularity of baseball in Brazil in the Culture Trip. Therefore we have included the age 20-24 Brazilian male population obtained from the UN into our MLB eligible population. The United Nations Census does not have information after 2010. We therefore estimate that the 2015 MLB eligible population is half of the 2010 MLB eligible population. We expect that this underestimates the actual 2015 MLB eligible population because we have observed a constant increase in the overall MLB eligible population as time increases.

	year	population	cumulative population proportion
1	1880	4.440	0.013
2	1890	5.010	0.027
3	1900	5.580	0.043
4	1910	8.560	0.068
5	1920	8.930	0.093
6	1930	9.920	0.122
7	1940	11.130	0.154
8	1950	11.590	0.187
9	1960	18.420	0.240
10	1970	24.490	0.310
11	1980	33.930	0.407
12	1990	37.460	0.515
13	2000	60.660	0.689
14	2010	72.270	0.896
15	2015	36.140	1.000

Table 1: Eligible MLB population throughout the years. The first column indicates the year, the second column indicates the estimated eligible MLB population size (in millions), and the third column indicates the proportion of the eligible MLB population in row x that was eligbile at or before row x.

3 The greats

At a quick glance of this population dataset, one can see how small the proportion of the pre-1950 eligible MLB population actually was. To determine which players are the all-time greatest players, we consult four lists which reflect popular opinion, performance metrics, and expert opinion that purport to determine the greatest players. The first list is compiled by ranker, which is constructed entirely from popular opinion as determined by up and down votes. The second and third lists rank players by highest career WAR as calculated by baseball reference and fangraphs, respectively. The fourth list is a ranking from ESPN and is based on expert opinion and statistics.

The rankings for all four lists are given in Table 2. As an example of the information contained in Table 2 consider the greatest players of all time according to ESPN displayed in the fourth column. We see that 5 players that started their careers before 1950 are in the top 10 all time and 11 players that started their careers before 1950 are in the top 25 all time. When the eligible MLB population is considered, it appears that the players from the earlier eras are overrepresented in this particular list. All of the lists that we consider exhibit this same bias towards the older players. In the next Section we provide statistical evidence supporting this position.

4 Statistical evidence

We now provide evidence that there are too many players that started their careers before 1950 included in top 10 and top 25 lists displayed in Table 2. We require two assumptions for the validity of our calculations which we will explore in detail in the next Sections. These assumptions are:

- First, we assume that innate talent is uniformly distributed over the eligible MLB population over the different eras.
- Second, we assume that the outside competition to the MLB available by other sports leagues after 1950 is offset by the increased salary incentives received by MLB players after 1950.

With these assumptions in mind we calculate the probability that at least x people from each top 10 and top 25 list in Table 2 started their career before 1950 using the proportion depicted in Table 1. Consider the bWAR list for example. According to bWAR, we see that 6 of the top 10 players started their career before 1950. From Table 1 we see that the proportion of the MLB eligible population that played at or before 1950 was approximately 0.187. We then calculate the probability that one would expect to observe 6 or more individuals in a top 10 list from that time period where the chance of observing each individual is about 0.187. We calculate this probability using the Binomial distribution.We perform the same type of extreme event calculation for each top 10 and top 25 list depicted in Table 2. The results are provided in Table 3.

As an example of how to interpret the results of Table 3, continue with bWAR's top 10 list. Table 3

rank	Ranker	bWAR	fWAR	ESPN
1	**Babe Ruth**	**Babe Ruth**	**Babe Ruth**	**Babe Ruth**
2	**Ty Cobb**	**Cy Young**	Barry Bonds	Willie Mays
3	**Lou Gehrig**	**Walter Johnson**	Willie Mays	Barry Bonds
4	**Ted Williams**	Barry Bonds	**Ty Cobb**	**Ted Williams**
5	**Stan Musial**	Willie Mays	**Honus Wagner**	Hank Aaron
6	Willie Mays	**Ty Cobb**	Hank Aaron	**Ty Cobb**
7	Hank Aaron	Hank Aaron	Roger Clemens	Roger Clemens
8	Mickey Mantle	Roger Clemens	**Cy Young**	**Stan Musial**
9	**Rogers Hornsby**	**Tris Speaker**	**Tris Speaker**	Mickey Mantle
10	**Honus Wagner**	**Honus Wagner**	**Ted Williams**	**Honus Wagner**
11	**Cy Young**	**Stan Musial**	**Rogers Hornsby**	**Lou Gehrig**
12	**Walter Johnson**	**Rogers Hornsby**	**Stan Musial**	**Walter Johnson**
13	**Joe Dimaggio**	**Eddie Collins**	**Eddie Collins**	Greg Maddux
14	Sandy Koufax	**Ted Williams**	**Walter Johsnon**	Rickey Henderson
15	Ken Griffey Jr.	**Pete Alexander**	Greg Maddux	**Rogers Hornsby**
16	**Jimmie Foxx**	Alex Rodriguez	**Lou Gehrig**	Mike Schmidt
17	**Tris Speaker**	**Kid Nichols**	Alex Rodriguez	**Cy Young**
18	**Joe Jackson**	**Lou Gehrig**	Mickey Mantle	Joe Morgan
19	Mike Schmidt	Rickey Henderson	Randy Johnson	**Joe Dimaggio**
20	Nolan Ryan	Mickey Mantle	**Mel Ott**	Frank Robinson
21	**Christy Mathewson**	Tom Seaver	Nolan Ryan	Randy Johnson
22	Roberto Clemente	**Mel Ott**	Mike Schmidt	Tom Seaver
23	Albert Pujols	**Nap Lajoie**	Rickey Henderson	Alex Rodriguez
24	**Cap Anson**	Frank Robinson	Frank Robinson	**Tris Speaker**
25	Greg Maddux	Mike Schmidt	Burt Blyleven	Steve Carlton
pre-1950 in top 10	7 / 10	6 / 10	6 / 10	5 / 10
pre-1950 in top 25	15 / 25	15 / 25	12 / 25	11 / 25

Table 2: Lists of the top 25 greatest baseball players to ever play in the MLB according to Ranker.com (1st column), bWAR (2nd column), fWAR (3rd column), and ESPN (4th column). Players that started their career before 1950 are indicated in bold. The last two rows count the number of players that started their careers before 1950 in each of the top 10 and top 25 lists respectively.

	Ranker	bWAR	fWAR	ESPN
probability of extreme event in top 10 list	0.000562	0.00448	0.00448	0.0249
probability of extreme event in top 25 list	0.0000057	0.0000057	0.000826	0.00322
chance of extreme event in top 10 list	1 in 1780	1 in 223	1 in 223	1 in 40
chance of extreme event in top 25 list	1 in 174816	1 in 174816	1 in 1210	1 in 310

Table 3: The probability and chance (1 in 1/probability) of each extreme event calculation corresponding to the four lists in Table 2.

shows that the probability of observing 6 or more players that started their career at or before 1950 of the top 10 all time players, based on population dynamics, is about 0.00448 (a chance of 1 in 223). The same interpretation applies to the other cells of Table 3. The results provided in Table 3 present overwhelming evidence that players who started their careers before 1950 are overrepresented in top 10 and top 25 lists from the perspectives of fans, analytic assessment of performance, and experts' rankings. These findings are a testament to 1) how sparse the MLB eligible population was at and before 1950 relative to the MLB eligible population after 1950 and 2) how incorporating relevant population dynamics leads to completely different conclusions.

5 Assumptions and Sensitivity Analysis

The probabilities and chances displayed in Table 3 are valid under the two assumptions given in the previous Section. In the first of these assumptions we specify that innate talent is evenly dispersed across eras. We do not fully believe that this assumption holds because the distribution of innate talent has improved over time as the eligible MLB population has expanded as noted by Stephen J. Gould, Christina Kahrl at ESPN, and in Martin B. Schmidt and David J. Berri's work on concentration of baseball talent in the Journal of Sports Economics. This suggests that the probabilities displayed in Table 3 are conservative. If we had refined data on the talent of those that strived to play professional baseball then the calculations in Table 3 would be even more extreme.

The second assumption states that the pool of talent available in the MLB has not been diminished by other sports leagues because of increased salary incentives to play baseball. We note that the National Basketball Association (NBA) and the National Football League (NFL) started in 1946 and 1920 respectively with both sports greatly rising in popularity since the inception of their respective professional leagues. That being said, it is widely known that the increases in MLB player salaries have been substantial. For example, the 1967 census lists the median US household income as $7,200. The minimum MLB salary at that time was $6,000 as noted by the LA Times sports writer Bill Shaiken in a piece titled "A look at how Major League Baseball salaries have grown by more than 20,000% the last 50 years." In 2015 the minimum MLB salary places earners around the top 1% of US household income. Brent Radcliff's article "Baseball Greats Who Were Paid Like Benchwarmers" has plenty more surprising MLB salary figures. In summary, baseball players made far less than they do today relative to the general US population. Furthermore, it is unclear that one could consider playing professional baseball to be a lucrative career in the older eras.

These strikingly different salary figures offer evidence that while other professional leagues may have drawn from the eligible MLB talent pool, salary incentives have led to an increase in the overall quality of MLB players. This suggests that our second assumption is conservative, again, if we had refined data on the eligible MLB population then the calculations in Table 3 would be even more extreme.

However, we are not convinced that this is the case with any certainty, it may be that our second assumption suffers some modest violations. To account for this possibility we consider a sensitivity analysis applied to the findings in Table 3. For this sensitivity analysis we weight the decennial populations displayed in Table 1 to reflect the overall interest that the US population has had in the game of baseball over time. The four specific weighting schemes, denoted by w1, w2, w3, and w4, and their full descriptions are given in the Appendix. We briefly describe these weighting schemes. The first and second weighting schemes weigh populations with respect to overall interest in baseball. No information is given for pre-1940 baseball. As a result of this we weigh years before 1940 with more weight than years after 1940 and w2 places more weight on pre-1940 populations than w1. The third weighting scheme weighs populations based on favorite sport information. No information is given for pre-1940 baseball. As a result of this we give the highest weight observed at or after 1940 for all pre-1940 years. The fourth weighting scheme is an average of w2 and w3. All of these weighting schemes suggest that the eligible MLB population was more interested in reaching the MLB in earlier eras than in modern eras.

The results of weighting populations and then recalculating the probabilities and chances displayed in Table 3 are displayed in Table 4. The conclusions from weighting populations with respect to w1 and w2 in Table 4 are consistent with those in Table 3. It is very unlikely that such a sparsely populated time period could have produced so many historically great baseball players. The third weighting scheme presents some conflicting conclusions. When weighting populations with respect to w3 we see that popular opinion, bWAR, and fWAR overrepresent players who started their careers before 1950. However the same is not so for the ESPN lists and the fWAR lists, the probabilities depicted are low but are not that extreme. This gives skeptics to our approach some evidence in their favor. We will return to this finding later. The conclusions from weighting populations with respect to w4 is largely in alignment with those of the original data, with a possible exception being the conclusion drawn from ESPN's top 10 list. The overall finding of this sensitivity analysis is that weighting populations with respect to fan interest in baseball does not alter the overall conclusion of the original analysis. It is very unlikely that such a sparsely populated time period could have produced so many historically great baseball players.

weight		Ranker	bWAR	fWAR	ESPN
w1	probability of extreme event in top 10 list	0.00121	0.00839	0.00839	0.0406
	probability of extreme event in top 25 list	0.0000267	0.0000267	0.0025	0.00845
	chance of extreme event in top 10 list	1 in 824	1 in 119	1 in 119	1 in 25
	chance of extreme event in top 25 list	1 in 37519	1 in 37519	1 in 401	1 in 118
w2	probability of extreme event in top 10 list	0.0023	0.0141	0.0141	0.0604
	probability of extreme event in top 25 list	0.0000944	0.0000944	0.00608	0.0182
	chance of extreme event in top 10 list	1 in 434	1 in 71	1 in 71	1 in 17
	chance of extreme event in top 25 list	1 in 10595	1 in 10595	1 in 164	1 in 55
w3	probability of extreme event in top 10 list	0.0274	0.0989	0.0989	0.256
	probability of extreme event in top 25 list	0.0102	0.0102	0.133	0.24
	chance of extreme event in top 10 list	1 in 36	1 in 10	1 in 10	1 in 3.9
	chance of extreme event in top 25 list	1 in 98	1 in 98	1 in 7.5	1 in 4.2
w4	probability of extreme event in top 10 list	0.00584	0.0296	0.0296	0.106
	probability of extreme event in top 25 list	0.000575	0.000575	0.021	0.0524
	chance of extreme event in top 10 list	1 in 171	1 in 34	1 in 34	1 in 9.4
	chance of extreme event in top 25 list	1 in 1740	1 in 1740	1 in 48	1 in 19

Table 4: The probability and chance (1 in 1/probability, rounded) of each extreme event calculation corresponding to the four lists in Table 2 after the eligible MLB population in Table 1 is weighted according to the four weighting schemes that are described in the Appendix.

6 Critique of additional comparison methods

In the previous Sections we observe that popular opinion (i.e., nostalgia) and performance metrics are in conflict with the population dynamics that underlie the distribution of the eligible MLB talent pool. We now critique additional methods that compare players across eras.

6.1 Critique of versus your peers methods

There are several methods which are used to compare players across eras that do so by computing a baseline achievement threshold within one season and then compare players to that baseline. These methods then rank players by how far they stood above their peers, the greatest players were better than their peers by the largest amount. As noted in the Introduction, examples include wins above replacement as calculated by baseball reference (bWAR), wins above replacement as calculated by fangraphs (fWAR), adjusted OPS+, adjusted ERA+, and computing normal scores. This manner of player comparison is purely statistical and it ignores talent discrepancies that exist across seasons. We have shown that this is a fundamentally flawed methodological approach that can exhibit major biases in player comparisons as evidenced by career bWAR and fWAR. Adjusted OPS+ is a worse offender than bWAR or fWAR and adjusted ERA+ is right in line with ESPN rankings.

6.2 Critique of detrending

In this Section we describe and critique the methodology of Petersen et al. (2011) in the context of comparing players across eras. We will refer to their paper as PPS due to repeated mentions. As described in PPS, they detrend player statistics by normalizing achievements to seasonal averages, which they claim accounts for changes in relative player ability resulting from both exogenous and endogenous factors, such as talent dilution from expansion, equipment and training improvements, as well as performance enhancing drug usage. As an argument in favor of players from earlier time periods, talent dilution from expansion is terribly misunderstood by PPS. If anything, the talent pool was more diluted in the earlier eras of baseball because of a small relative eligible population size and the exclusion of entire populations of people on racial grounds. See Table 5 for the specifics. As an argument in favor of old time players, equipment and training improvements is not without fault because the same improvements are available to every competitor as well. PPS also fails to account for the effect of modern day filming as a way to gain insight on opponents' strengths and weakness in their training improvements assumption. PPS also fails to account for increases in salary compensation enjoyed by MLB players in more modern eras.

The principle assumptions about baseball used as justification for PPS detrending are head scratchers at best. The mathematics of PPS detrending is also questionable in the context of comparing baseball players across eras. PPS notes that the evolutionary nature of competition results in a non-stationary rate of success. As already noted, PPS detrends player statistics by normalizing achievements to seasonal averages. The normalization goes as follows: Suppose a player hits 40 homeruns in a given season and that the league average prowess for homerun hitting in that season is 10 homeruns. If the historical average prowess for homerun hitting is 5 homeruns then our player's detrended homerun count in that particular season is $40 \times (5/10) = 20$. In general, the detrending formula is $Y \times (\text{historic prowess}/\text{league prowess})$ where Y is individual prowess for a particular player in a given season. Fundamentally different approaches for detrending are advocated in authoritative textbooks such as Introduction to Time Series and Forecasting, by Peter J. Brockwell and Richard A. Davis, Time Series Analysis and Its Applications With R Examples, by Robert H. Shumway and David S. Stoffer, and Time Series Analysis and Forecasting by Example, by Søren

year	eligible pop.	number of teams	roster size	eligible pop. per roster spot
1880	4.44	8	15	37
1890	5.01	8	15	41.7
1900	5.58	8	15	46.5
1910	8.56	16	25	21.4
1920	8.93	16	25	22.3
1930	9.92	16	25	24.8
1940	11.13	16	25	27.8
1950	11.59	16	25	29
1960	18.42	16	25	46.1
1970	24.49	24	25	40.8
1980	33.93	26	25	52.2
1990	37.46	26	25	57.6
2000	60.66	30	25	80.9
2010	72.27	30	25	96.4

Table 5: Relative talent dilution when considering the eligible MLB population sizes at select time periods. Eligible population totals are in millions in column 2 and are in thousands in column 5.

Bisgaard and Murat Kulahci, and all of the authors conversations with statisticians.

With all of this being said, we see PPS detrending as more of an inflationary metric of relative prowesses and not an era adjustment. We agree with Petersen when he says, "detrending corresponds to removing the inflationary factor, so we could compare two items like the cost of a candy bar in 1920 to the cost of a candy bar in 2000. In this case, we compare Babe Ruth's home runs–the ability of someone to get a home run then versus now–and you see Babe Ruth actually hit a lot of home runs on this relative basis," in a BU Today piece written by Rachel Johnson. However, having higher prowess relative to your peers, hitting more runs in this case, is not indicative of a player's prowess with respect to peers from fundamentally different eras. Additionally, failing to account for the fact that baseball was segregated before 1947 does not make the statistics fairer despite the claim that PPS detrending is a way to do as such.

6.3 Critique of era bridging

Berry et al. (1999) claim that their era bridging technique accounts for talent discrepancies across eras. However, they do not explicitly parameterize this in their hierarchical models. They also conclude that "globalization has been less pronounced in the MLB (relative to other sports), where players are drawn mainly from the United States and other countries in the Americas. Baseball has remained fairly stable within the United States, where it has been an important part of the culture for more than a century,“ (Berry et al., 1999). This rationale ignores the segregation of baseball before 1947, increases in the eligible MLB population relative to available roster spots, and increases in the average overall talent of the eligible MLB population. They claim that they capture the changing pool through use of separate distributions for each decade which allows them to study the changing distribution of players in sports over time (Berry et al., 1999). This type of adjustment is calculated on the achievement itself, not the traits of the individuals producing the achievement. Therefore it does not study the characteristics of a changing talent pool. We can see this in Berry et al. (1999, panel (c) of Figure 7). In this figure we see that their model predicts that a .300 hitter in 1996 will have a lower than .300 average for several seasons from 1900-1920. This conflicts with the well established notion that the talent of baseball players has improved over time.

We can see that their methodology does not incorporate changing population dynamics through examining their Berry et al. (1999, Table 9). In this table they find that 6 of the 10 best hitters for average started their career before 1950 and 10 of the 25 best hitters for average started their career before 1950. Their paper came out in 1999 so we recompute the chances of these events where the eligible MLB population ends at 1999 (we approximate this by taking the population counts in the first 12 rows of Table 1 and then multiplying the population count in 2000 by 0.9). We calculate the chance that one would expect to observe 6 or more individuals in a top 10 list who started their career before 1950 as 1 in 30. We calculate the chance that one would expect to observe 10 or more individuals in a top 25 list who started their career before 1950 as 1 in 7.7. These chances are not as extreme as those in Table 3 but they still correspond to events that are highly unlikely.

6.4 Critique of our testing procedure

Our testing procedure is valid under two assumptions made in Section 4. A sensitivity analysis is given in Section 5 that considers how our findings change under violations of our assumptions. The results of the sensitivity analysis are largely consistent with our original analysis, with an exception being when we weigh populations with respect to a weighting scheme which states that the eligible MLB population were those who listed baseball as their favorite sport.

year	original data	w1	w2	w3	w4
1880	37	18.5	22.2	14.1	18.1
1890	41.7	20.9	25.1	15.9	20.5
1900	46.5	23.2	27.9	17.7	22.8
1910	21.4	10.7	12.8	8.1	10.5
1920	22.3	11.2	13.4	8.5	10.9
1930	24.8	12.4	14.9	9.4	12.2
1940	27.8	11.1	11.1	9.7	10.4
1950	29	11.6	11.6	11	11.3
1960	46.1	18.4	18.4	15.7	17
1970	40.8	16.3	16.3	11.4	13.9
1980	52.2	20.9	20.9	8.4	14.6
1990	57.6	23.1	23.1	9.2	16.1
2000	80.9	32.4	32.4	10.5	21.4
2010	96.4	38.5	38.5	11.6	25.1

Table 6: Relative talent dilution when considering the weighted eligible MLB population sizes. Numbers are in thousands of people per roster spot.

7 Discussion

Our methodology holds up well when we weight our original data in ways that reflect relative disbelief in our assumptions. Moreover, we expect that the conclusions drawn from weighting with respect to our third weighting scheme are extremely unlikely. To see this we investigate talent dilution with respect to the four weighted populations as depicted in Table 6. From the w3 column of Table 6, we see that talent was least diluted at and before 1900. If that were truly the case then we would expect less variability and outliers in achievements as described by Stephen J. Gould in numerous books, articles, and YouTube videos. The baseball reference leader boards for these years indicate the exact opposite. In particular, see the season batting average leaders. This suggests that the talent pool in the MLB was diluted during that time period.

PPS does not offer any insight on how their method fairs when their underlying assumptions are called into question. As argued in Section 6.1, methodology which compares players across eras by examining who stood the farthest from their peers is fundamentally flawed. Era bridging offers a step in the right direction but the methodology does not explicitly account for a changing talent pool.

The MLB players from the old eras of baseball receive significant attention and praise as a result of their statistical achievements and their mythical lore. We argue that these players are collectively overrepresented in rankings of the greatest players in the history of the MLB. Our statistical methodology provides overwhelming evidence in favor of this argument. We conclude that the superior statistical accomplishments achieved by players that started their careers before 1950 are a reflection of our inability to properly compare talent across eras. It is highly unlikely that athletes from such a scarcely populated era of available baseball talent could represent top 10 and top 25 lists so abundantly.

Appendix: Conservative Weighting Schemes

The weights discussed in Section 5 are given in the table below. These weights reflect Gallup polling data on the subject of fan interest in baseball. Polling data does not exist for all time periods. Specifically, there are no records before 1940. The first two weighting schemes reflect overall fan interest in baseball. It is apparent from the Gallup dataset that the proportion of Americans sampled who are baseball fans has been around 40% from 1940 on. The first weighting scheme (w1) places a 0.50 weight on fan interest with the assumption that more Americans from those time periods favored, and played, baseball due to lack of competition from other sports. The second weighting scheme (w2) places even stronger weights on pre-1940 fan interest. The third weighting scheme (w3) reflects the proportion of Americans sampled that list baseball as their favorite sport. Again, we place a higher weight for pre-1940 fan interest. The fourth weighting scheme (w4) is the average of w2 and w3.

All four weighting schemes serve as an educated proxy for the eligible MLB population thought to strive towards a career in professional baseball. The weighting schemes w2, w3, and w4 are especially conservative. We do not expect fan interest in pre-1940s baseball to be as high as our weighting schemes due to lack of media exposure and salary compensation. The appropriateness of w3 is particularly questionable since individuals play baseball even if it is not the individual's favorite sport. It should be noted that these weights are obtained from survey data from the United States only and are applied, at equal value, to the other countries. Therefore our weighting schemes are biased but rich survey data on the topic of fan interest in baseball is unavailable from from the other countries.

References

Berry, S. M., Reese, C. S., and Larkey, P. D. (1999), "Bridging Different Eras in Sports," *Journal of the American Statistical Association*, **94**, 447, 661–676.

Petersen, A. M., Penner, O., Stanley, H. E. (2011), "Methods for detrending success metrics to account for inflationary and deflationary factors," *The European Physical Journal B*, **79**, 67–78.

	year	w1	w2	w3	w4
1	1880	0.50	0.60	0.38	0.49
2	1890	0.50	0.60	0.38	0.49
3	1900	0.50	0.60	0.38	0.49
4	1910	0.50	0.60	0.38	0.49
5	1920	0.50	0.60	0.38	0.49
6	1930	0.50	0.60	0.38	0.49
7	1940	0.40	0.40	0.35	0.38
8	1950	0.40	0.40	0.38	0.39
9	1960	0.40	0.40	0.34	0.37
10	1970	0.40	0.40	0.28	0.34
11	1980	0.40	0.40	0.16	0.28
12	1990	0.40	0.40	0.16	0.28
13	2000	0.40	0.40	0.13	0.27
14	2010	0.40	0.40	0.12	0.26
15	2015	0.40	0.40	0.10	0.25

Table 7: Weighting schemes.

Schmidt, M. B. and Berri, D. J. (2005), "Concentration of Playing Talent: Evolution in Major League Baseball," *Journal of Sports Economics*, **6**, 412–419.