

Baseball players with the initial “K” do not strike out more often

B.D. McCullough* and Thomas P. McWilliams

Department of Decision Sciences, Drexel University, Philadelphia, PA 19104, USA

(Received 28 April 2008; final version received 11 March 2009)

It has been claimed that baseball players whose first or last name begins with the letter K have a tendency to strike out more than players whose initials do not contain the letter K. This “result” was achieved by a naive application of statistical methods. We show that this result is a spurious statistical artifact that can be reversed by the use of only slightly less naive statistical methods. We also show that other letters have larger and/or more significant effects than the letter K. Finally, we show that the original study applied the wrong statistical test and tested the hypothesis incorrectly. When these errors are corrected, most of the letters of the alphabet have a statistically significant strikeout effect.

Keywords: name-letter effect; spurious correlation

1. Introduction

In a recent article that has been widely reported in the popular media, Nelson and Simmons [4] (hereafter “NS”) reference the so-called “name-letter effect” [5] and claim that baseball players with the first or last initial K tend to strike out more often than other players. A less selective and more scrupulous reporting of results (using their incorrect method of analysis) would have said, “Baseball players with the initials K, D, and P tend to strike out more often than players with other initials, and players with the initials E, F, G, H, and O tend to strike out less than other players.” However, when the data are analyzed, *all* the initials except C, M, R, U, and V are statistically significant! There are other problematic issues with their article that we shall address.

NS theorize that baseball players whose first and/or last names begin with a “K” strike out more than other baseball players (for brevity we shall just refer to them as players whose name begins with K or even K-players). While strikeouts are bad and players would prefer to avoid them, they contend that players with an initial “K” (which signifies “strikeout” on a score sheet), presumably because of an affinity for their own initial, would not be so averse to strikeouts as other players. This decreased aversion (i.e. self-sabotage) should translate into more strikeouts for K-players (i.e. self-sabotage). To test their theory, they analyze data for the years 1913–2006. For each major league player with at least 100 lifetime plate appearances (they report this number to

*Corresponding author. Email: bdmccullough@drexel.edu

be $N = 6397$), they define the strikeout percentage to be 100 times the number of strikeouts (SO) divided by the total number of plate appearances (the sum of at-bats (AB) and bases-on-balls (BB)). Their article describes their

Main Result: “Across more than 90 years of professional baseball, batters whose names began with a K struck out at a higher rate (in 18.8% of their plate appearances) than the remaining batters (17.2%), $t(6395) = 3.08$, $p = 0.002$.” (p. 1107)

They also adduce additional results to buttress their theory [4, p. 1007]:

- [AR1] K-players strike out more often “when we controlled for the average year in which each athlete played ($p < 0.015$).”
- [AR1A] “When we controlled for average year of play (and excluded initials associated with fewer than 5 Major League players – e.g. U as a first initial), K was both the first initial and the last initial associated with the highest strikeout rate.” (This claim is subordinate to AR1.)
- [AR2] Controlling for “American or foreign born also showed that batters with the initial K were reliably more likely to strike out than other players were ($p = 0.023$).”
- [AR3] K-players struck out more often “when we controlled for country of origin with a dummy variable for each of the 52 countries represented in the sample ($p = 0.045$).”

Neither their main result nor any of their additional results withstands scrutiny. First, as we show in Sections 3 and 4, there is nothing special about the letter K. No matter how the data are analyzed, K is not unique and many letters are significant, yet NS neglect to mention this. Second, their hypothesis test implementation is incorrect. All the p -values they present are for two-sided tests. Their hypothesis, however, is that K-players strike out *more* than other players, which necessarily implies a one-sided test. Therefore, NS have employed two-sided tests when they should have employed one-sided tests. This is not a trivial mistake for a professor to make.

Second, as we show in Section 3, NS choose the wrong unit of observation. They use the player rather than the plate appearance as the unit of observation, which leads them incorrectly to apply a two-sample test of means rather than a two-sample test of proportions. Third, as we show in Section 4, even if the player is the correct unit of observation, they ignore massive heteroscedasticity that, when corrected, reverses their results. Fourth, they neglected to consider whether the effect they found might be spurious; it is, as we show in Section 5. All this requires that we can first replicate the results in the NS paper, which we do in the next section.

2. Replicating the NS results

We first seek to reproduce the results in the Main Result. NS use the database at www.baseball-databank.org, and a simple SQL command on the Master and Batting tables produces the data set of 6397 players. Statistics on these players are presented in Table 1, which reproduce NS’s 18.8% for K-players and 17.7% for non-K players.

Plugging the mean, standard deviation, and sample size into the usual two-sample test of means with equal variances formula yields $t = 3.0753$, $df = 6395$, with a two-sided p -value = 0.0021 which reproduces NS’s main result. Thus we are confident that we are using the correct data and have analyzed it the same way they did. Since NS would neither supply the data set they used, nor the SQL code that they used to create their data set, nor the code that they used to analyze the data set, we had to spend a substantial amount of time reverse-engineering their results. In the course of this reverse-engineering, we discovered several errors that we describe in the Appendix.¹

Due to space constraints, NS did not describe the methods behind their additional statistical evidence (AR1, AR2, and AR3, above) by which they “controlled for” variables in the above.

Table 1. Statistics for baseball players, 1913–2006.

	K players	Non-K players
n	377	6020
\sum playso	90,263	1,380,367
\sum playABBB	657,812	10,704,712
\sum playso / \sum playABBB	0.1372	0.1289
Mean of soperc	0.1881	0.1772
Standard deviation of soperc	0.10643	0.10006

Notes: n , number of players; playso, a player's number of strikeouts; playABBB, a player's plate appearances (at-bats plus bases-on-balls); soperc, playso/playabb, an individual player's strikeout percentage.

As there is no obvious analog to the two-sample test for means by which one might control for these variables, we had to ask NS how they did this.

They informed us that they used the analysis of covariance (ANCOVA) method, which involves a regression model that includes, as independent variables, a dummy “K-player” variable along with one or more quantitative variables called covariates (see, e.g. [3, p. 917]). To understand the underlying logic, first note that an equal-variances two-sample t -test to see if the strikeout rate is higher for K-players can be implemented via the usual formula or by running a simple linear regression with strikeout percentage as the dependent variable and a K-player dummy variable as the independent variable. The t -statistic for the dummy variable will be the same as that obtained using the two-sample t -statistic formula, and the p -values will, therefore, be identical.

The regression approach to testing for a difference in strikeout rates allows expansion of the model to account or “control” for other variables (such as average year) that might help to explain variation in the strikeout percentage. If these additional variables, called covariates, are indeed related to the strikeout percentage then their inclusion can serve to reduce the variance of the error terms in the regression model, leading to a more precise analysis.

3. Discussion of NS methodology

Via selective reporting, NS convey the (mis)impression that the letter K is somehow special, yet when the method used by NS is applied to all the letters of the alphabet, it turns out that K is hardly unique, as shown in Table 2. For example, K, D and P have statistically significant positive differences, indicating that these players strike out more often than others. E, F, G, H and O have statistically significant negative differences, indicating that they strike out less often than other players. Yet NS make no mention of any of these effects, save the letter K. Nor is K the letter with the biggest difference. The effect for the letter O is -0.0168 , which is greater (in absolute value) than the effect for the letter K (0.0164), yet NS did not mention this pronounced effect. Further, if the name-letter effect is real and exists in baseball, why should it exist only for the letter K and strikeouts? Why do NS neglect to consider also S for singles, H for homeruns, D for doubles, O for outs, E for errors, and so on?

NS do not actually say that they tested all the letters for this simple strikeout effect, but they admit to testing all the letters in some cases that were more difficult to program (e.g. their “additional result” ARIA), of which they write, “K was both the first initial and last initial associated with the highest strikeout rate.” It strains credulity to think that they did not test all the letters in the simple case, so it is reasonable to suppose that they were aware that O-players are even more strikeout averse than the much-touted K-players are strikeout prone. Surely if the latter was worth reporting, the former was, too; one has to wonder why NS chose not to mention this. Perhaps

Table 2. Two-sample test of means for each letter.

	Difference
A	0.0018
B	0.0013
C	-0.0023
D	0.0135*
E	-0.0132*
F	-0.0121*
G	-0.0114*
H	-0.0098*
I	-0.0153
J	0.0037
K	0.0164*
L	-0.0033
M	-0.0010
N	0.0062
O	-0.0168*
P	0.0114*
Q	0.0049
R	0.0042
S	-0.0011
T	0.0055
U	-0.0240
V	-0.0050
W	-0.0006
X	-na-
Y	-0.0237
Z	-0.0097

Notes: *Significance at the 5% level.

because it undermines their theory of self-sabotage: what form of anti-self sabotage accounts for this “statistically significant” difference for O-players?

The theory promoted by NS led them to predict “that players whose first or last names begin with K would show an increased tendency to strike out.” Note that the reference is to “players” which is plural and collective, not to an individual player. The hypothesis, as stated, compares all K-players with all non-K-players. To effect this comparison, one should compute a strikeout percentage for the group of K-players and the group of non-K-players, and compare these two percentages. To measure the propensity of each group to strike out, NS calculated 18.8% for K-players and 17.2% for non-K players. However, they calculated the wrong measure. This raises the issue of the correct unit of observation: is it the player or the plate appearance? The choice is not innocuous, for choosing the player leads to a two-sample test of means, while choosing the plate appearance leads to a two-sample test of proportions. NS chose the player when, in fact, the correct choice is the plate appearance.

To illustrate the point in the present case, consider the three “K players” with the lowest and the three with the highest strikeout percentages (Table 3): what is the proportion of strikeouts for these six players? NS calculated it this way:

$$\text{NS method} = \frac{0.0392 + 0.0401 + 0.0423 + 0.5629 + 0.5426 + 0.5517}{6} = 0.2965,$$

which is incorrect because it employs an equal-weighting strategy that does not take into account the differing number of at-bats for each player, and consequently will not give the proportion of

Table 3. K players with lowest and highest strikeout percentage.

Name	ABBB	SO	soperc
Kell, George	7323	287	0.0392
Killefer, Bill	2741	110	0.0401
Koenig, Mark	4493	190	0.0423
Kaiserling, George	186	98	0.5629
Kelley, Dick	129	80	0.5426
Kilkenny, Mike	116	64	0.5517

strikeouts for the group of players.² The correct method is given by:

$$\text{Correct method} = \frac{287 + 110 + 190 + 186 + 80 + 64}{7323 + 2741 + 4493 + 186 + 129 + 116} = 0.06118.$$

Clearly the correct answer is 0.06118 and not 0.2965, though NS would have us believe that the latter number is correct. To see that it is not, simply multiply the group’s number of plate appearances, 14988, by 0.2965, to obtain 4444, which does not equal the group’s number of strikeouts, 917. However, 14,988 times 0.06118 equals 917. To compute the proportion of strikeouts for the entire group, it is misleading to pretend that they all had the same number of plate appearances, as the NS method does. Indeed, the correct method is the one used to calculate team batting averages and league batting averages.

Additionally, since NS test hypotheses, we must consider the standard errors that they calculate. The two-sample test of means is meant to be applied to i.i.d. data, yet due to the markedly differing number of at-bats for individual players (ranging from 100 to 15,619), the sequence of player strikeout percentages is heteroscedastic. Striking out is a yes–no proposition; the sequence of plate appearances gives rise to a binomial distribution, for which the variance is $p(1 - p)/n$. Using each player’s estimated strikeout rate as p and number of plate appearances as n , we calculated estimated variances ranging from 1.768E-6 to 0.00246, for a ratio of maximum to minimum variance of roughly 1400:1. Massive heteroscedasticity is clearly present in the data and must be adjusted for in any credible analysis.

Hypothesis testing requires accurate standard errors, which cannot be obtained via ordinary least squares (OLS) in the presence of heteroscedasticity, and a correction is in order. The usual solution is to weight by the inverse of the variance (i.e. weighting each batter by the number of plate appearances). One way to do this is simply to use the two-sample test of proportions. Accordingly, if π_k is the proportion of times that K-players strike out and similarly π_{nk} corresponds to non-K-players, then the relevant null and alternative hypotheses are

$$H_0: \pi_k - \pi_{nk} \geq 0,$$

$$H_A: \pi_k - \pi_{nk} < 0.$$

These hypotheses contrast sharply with the hypotheses tested by NS. Thus, even if the player and not the plate appearance is the unit of observation, due to the necessity of treating heteroscedasticity, the two-sample test of means employed by NS is incorrect and the two-sample test of proportions method is correct.

4. Discussion of NS additional results

In the case of the ANCOVA regressions (AR1, AR2, and AR3) that NS used to support their hypothesis, the proper statistical course of action is to correct for the heteroscedasticity. We will

do this by performing weighted least squares (WLS) regression analyses and will compare the results with those obtained via the OLS approach used by NS. We model the variance function as $p(1 - p)/n$, as discussed in Section 3, where p is the strikeout proportion and n is the number of plate appearances. The inverse of the variance function is used as the weight function (see, e.g. [3, p. 425]).

The ANCOVA equation to be estimated regresses a player’s strikeout percentage on an intercept, a dummy variable corresponding to whether the player is a K-player, the average year in which the player played, a dummy for whether the player is born in the USA, and 42 dummy variables corresponding to the player’s birth country (results for the 42 are not presented, but summarized). The results for AR1–AR3 are given below, where AR1 is the usual linear regression corresponding to AR1, and AR1-W is the WLS regression.

The first thing to note about Table 4 is that K-player is insignificant in all the WLS regressions. Consequently, none of the additional results adduced by NS is valid. Note also that the USA dummy, significant in the NS regression that incorrectly assumes homoscedastic errors, becomes insignificant in the heteroscedasticity-consistent WLS regression. Surprisingly, not a single one of the 42 country dummies in AR3 was significant. To claim that one has “controlled for country of origin” when all the country dummies are insignificant is specious, as best. In the case of AR3-W, a single one of the 41 country dummies is significant at the 5% level.

Even though NS used the wrong approach (ANCOVA regression without heteroscedasticity correction), if they had applied it correctly by checking for heteroscedasticity and then correcting for heteroscedasticity, they would have found that K-players do not strike out more often.

We know that applying a two-sample test of means or ANCOVA regression are invalid methods for testing the hypothesis of interest. The correct way to test $H_0: \pi_k - \pi_{nk} \geq 0$ against $H_A: \pi_k - \pi_{nk} < 0$ is to calculate

$$\hat{\pi}_k = \frac{90263}{657812} = 0.1372, \quad \hat{\pi}_{nk} = \frac{1380367}{10704712} = 0.1289$$

and perform the usual two-sample z -test of proportions, taking care to perform a one-sided test yielding $z = 19.5$ (p -value = 0) which clearly rejects the null hypothesis. As will be made clear, this “significant” result is due to the well-known phenomenon that everything is significant when the sample size is sufficiently large (see, e.g. [2]). In such a situation, it becomes critical to be aware of the difference between “statistical significance” and “practical significance.”³ Note that the method of proportions automatically weights player’s performance by her number of

Table 4. OLS and weighted (-W) least squares regressions of AR1, AR2, and AR3.

Variable	AR1	AR1-W	AR2	AR2-W	AR3	AR3-W
Intercept	-1.96	-2.15	-2.08	-2.18	-2.08	-2.19
p -Value	0	0	0	0	0	0
K-player	0.013	0.005	0.011	0.004	0.010	0.004
p -Value	0.012	0.083	0.028	0.117	0.044	0.138
Avg. year	0.0011	0.0012	0.0011	0.0012	0.0012	0.0012
p -Value	0	0	0	0	0	0
USA dummy	-	-	0.018	0.004	-	-
p -Value	-	-	5E - 6	0.055	-	-
Birth country	-	-	-	-	n.r.	n.r.
p -Value	-	-	-	-	All > 0.05	One < 0.05
R^2	0.09768	0.307	0.1006	0.3074	0.1105	0.3162

Notes: n.r.: 42 birth country dummy variable coefficients not reported.

plate appearances, so it is immune to the problem that invalidates the two-sample test of means approach. Thus, it might appear that K-players do strike out more often than other players, but such a conclusion would be premature.

We now show that such a significant result is to be expected, by testing *all* the letters of the alphabet. Since we are only concerned with whether there are differences and not hypothesizing that one type of player strikes out more (or less) than others, our tests of proportions (which is the correct method) are two-sided. The procedure is as follows:

- (1) For the letter A, split players into two groups: those whose first or last names begin with A, versus players who do not meet this criterion.
- (2) Perform a two-sided two-sample z -test for a difference in proportions who strike out for the two groups.
- (3) Repeat this for all 26 letters.

Since 26 tests are being performed, control the family or group significance level (Type I error probability) at 0.05. This is done using the Bonferroni procedure: perform each individual test at significance level $0.05/26 = 0.001923$. The corresponding critical z -value for the individual two-sided tests is ± 3.10 . Results are presented in Figure 1.

Note that 21 of the 26 letters are significant, reinforcing our contention that “everything is significant” and that NS found nothing worth noting when they concluded that K-players strike out more, because the letters A, C, D, J, N, P, Q, T, V, and X also strike out more, while B, E, F, G, H, I, L, M, O, S, W, Y, and Z strike out less. The only thing worth remarking is that the letters C, M, R, U, and V are *not* statistically significant! This is consistent with running a logistic regression (results not reported) with dummy variables for the 26 letters of the alphabet. At the 5% level, the only insignificant letters were A, B, C, and X; in other words, almost every letter was significant.

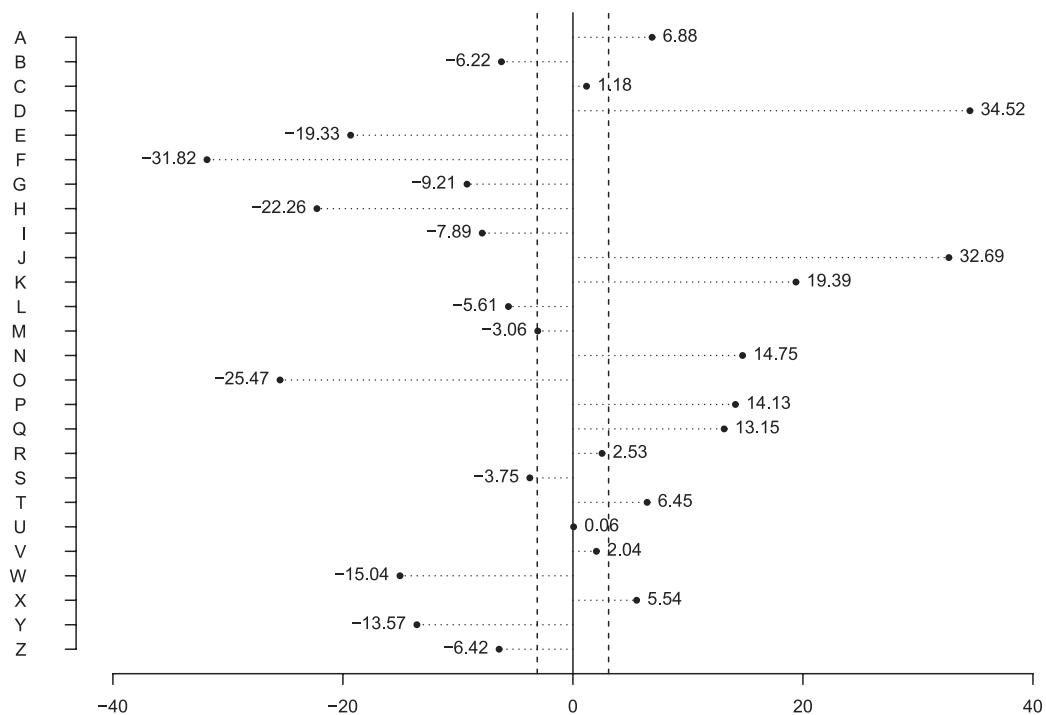


Figure 1. z -Scores for Bonferroni test of all 26 letters. Vertical dashed lines are critical values, ± 3.102 .

5. Is the NS result spurious?

It is all well and good to form an hypothesis about the letter K and to test that hypothesis. However, if the researcher is going to test K against other letters, as NS did, the proper statistical framework is to first test the null hypothesis: “Do all the letters have the same strikeout rate?” and then inquire which (if any) differences are significant. Such an approach is problematic in the present case because the columns are not independent. The player “Frank Howard” will appear twice, both as an F-player and an H-player. This correlation will bias the test toward rejecting the null. If players happen to be randomly high or low, the fact that they are counted twice will artificially inflate the *F*-statistic.

How might we check for spuriousness? If the effect is real, it will be present in all subsamples, though perhaps the sample size will not be large enough to detect the effect if there are too many subsamples. For example, if we divide the data set into two parts, 1913–1959 and 1960–2006, we find the summary statistics for these two subperiods presented in Table 5.

If NS had applied their (incorrect) method of testing two-sample means to Table 5, they would have found that in the first half of the sample, K-players struck out more frequently than non-K players, $t(2975) = 3.842$, $p < 0.01$, but in the second half there was no statistically significant difference between 0.2045 and 0.2000, $t(3678) = 0.66$, $p = 0.51$. Their “result” that they claim applies to 1913–2006 appears to be driven by what happened in the first half of the century. One wonders how NS might change their theory to account for these results; did K-players in the second half of the century learn to stop sabotaging themselves?

Dividing the sample into two subperiods raises another issue: nonstationarity (we thank a referee for pointing this out). As can be seen from Table 5, players were much less prone to strike out in the first half of the sample, $(20676 + 381745)/(209683 + 4141439) = 0.092$, compared with the second half of the sample, $(69466 + 996736)/(447746 + 6556343) = 0.152$. This undermines the implicit assumption underlying the NS analysis that there is a constant strikeout proportion throughout the entire sample. A full-blown analysis of these data should take into account the fact that the strikeout proportion changes over time.

As mentioned, we would not want to break this down into too many subsamples, because the subsamples might not be large enough to discern a true effect. To see this, consider a simple Monte Carlo exercise.

Table 5. Statistics for baseball players.

	K players	Non-K players
1913–1959		
<i>n</i>	161	2816
\sum playso	20676	381745
\sum playABBB	209683	4141439
\sum playso/ \sum playABBB	0.0968	0.0922
Mean of soperc	0.1629	0.1357
Standard deviation of soperc	0.10760	0.08608
1960–2006		
<i>n</i>	233	3447
\sum playso	69466	996736
\sum playABBB	447746	6556343
\sum playso/ \sum playABBB	0.1551	0.1520
Mean of soperc	0.2045	0.2000
Standard deviation of soperc	0.10483	0.10049

Table 6. Hypothesis testing on subsamples testing $H_0:\mu = 0$ when the data are $N(3, 100^2)$.

Subsample	<i>t</i> -Statistics	<i>p</i> -Value
1	1.7807798	0.0376261
2	1.7541884	0.0398524
3	1.6404676	0.0506114
4	0.3560920	0.3609234
5	0.6546494	0.2564221
6	1.1367361	0.1279606
7	1.1109166	0.1334358
8	− 1.176217	0.8801058
9	0.6154836	0.2691878
10	− 1.730597	0.9580838

Generate a 10,000 random normals with mean 3 and standard deviation 100, and test the null hypothesis that the mean is greater than zero against the alternative that it is not. For our particular random numbers, the result is $t = 1.9446$, $p = 0.02592$. If, however, we split the sample into 10 subsamples of 1000 each, we obtain Table 6.

Even though the population mean is 3, generally the null that the mean is zero is not rejected. Note that only three of the subsamples are significant, and a pair of *t*-statistics even have negative signs. The effect would become even more pronounced if we were to allow the sample sizes of the subsamples to vary. This example has to do with not finding an effect that is present. What about finding an effect that is not present?

To find an effect that is not present, one could start with a large sample drawn from a common distribution, then arbitrarily draw subsamples, and calculate the mean of each. Take a subsample whose sample mean is above the population mean, and compare it with the remaining observations. The sample mean of the remaining observations will almost surely be below the population mean, so the stage is set to find a “significant” difference where one does not really exist.

In this spirit, then, we ask: suppose that 6397 players are drawn from a common distribution that has a mean strikeout percentage of 17.2%, and these players are randomly assigned to 25 categories denoted by the letters A–Z, where the sample sizes of the subsamples conform to the sample sizes of the subsamples analyzed by NS. What might we expect to discover? In a small simulation⁴ we used a scaled and centered chi-square distribution with two degrees of freedom to represent the outliers and right-skewed nature of the real data, and found that eight letters were “significant”: H, L, M, O, P, S, V, and W. Specifically, the letters L, O, S, and V all have positive significant *t*-statistics, while the letters H, M, P, and W have negative significant *t*-statistics. This simulation is similar to the situation analyzed by NS, and we have found that there are “statistically significant” letters even though we know that all the letters have the same true mean strikeout percentage. The “effect” that NS discovered is spurious.

6. Conclusion

NS [4] theorized that the so-called “name-letter effect” induced baseball players whose first or last name begins with the letter K to strike out more often than other players. Aside from selectively ignoring other implications of the name-letter effect (Do players with the initial H hit more homeruns?) and selectively reporting results (they report that K is the most strikeout prone letter *after* controlling for average year of play, but neglect to mention that without controlling it is only one of nine statistically significant letters and is neither the “most significant” nor the one with the largest effect), NS apply the wrong statistical test (two-sample test of means with equal weights instead of two-sample test of proportions), test the hypothesis incorrectly (report a

two-sided p -value for an obviously one-sided test), and ignore the massive heteroscedasticity that invalidates their hypothesis testing. We also show that even if the two-sample test of means was appropriate for this situation, that any significant results would be spurious statistical artifacts. When we use the correct method to test the hypothesis, we find that most letters are statistically significant with respect to strikeouts. This, too, is a meaningless result, driven by the fact that when the sample size is large enough, everything is significant.

Pelham et al. [6] collected some data, analyzed it, and claimed that the data supported the name-letter effect. Gallucci [1] reanalyzed the Pelham et al. data and found that the data did not support the name-letter effect. We have done likewise here. NS made other claims in their paper. These claims are the subject of current research.

Acknowledgements

We thank an anonymous referee for many valuable suggestions.

Notes

1. We thank Simmons for verifying the errors that we uncovered.
2. Reweighting the strikeout percentages can yield a reliable answer, but NS do not do this.
3. While we are aware of no studies that directly address this point, anecdotal evidence suggests that the difference is not practically significant. In 500 plate appearances, the difference between 0.1372 and 0.1289 amounts to four strikeouts per season. Quoting Kevin Costner as Crash Davis in the movie "Bull Durham" is instructive: "Twenty-five hits a year in 500 at bats is 50 points . . . That's about 25 weeks. You get one extra flare a week – just one – a gork, a ground ball with eyes, a dying quail – just one more dying quail a week and you're in Yankee Stadium." It appears, then, that 25 of 500 is practically significant, and 4 out of 500 does not come close to this standard.
4. As noted previously, the columns are correlated, which leads to a greater number of significant letters than if the columns were uncorrelated. For our simulation, we could induce a correlation in our columns, but reproducing the exact correlations as are present in the original data would be time-consuming. To achieve the same effect, then, we simply chose a simulation run that has a similar number of significant letters.

References

- [1] M. Gallucci, *I sell seashells by the seashore and my name is Jack: Comment on Pelham, Mirenberg, and Jones (2002)*, *J. Pers. Soc. Psychol.* 85(5) (2003), pp. 789–799.
- [2] J.B. Kadane, *Testing precise hypotheses: Comment*, *Statist. Sci.* 2(3) (1987), pp. 347–348.
- [3] M.H. Kutner, C.J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, McGraw-Hill, New York, 2005.
- [4] L.D. Nelson and J.P. Simmons, *Moniker maladies: When names sabotage success*, *Psychol. Sci.* 18(12) (2007), pp. 1106–1112.
- [5] J.M. Nuttin, *Affective consequences of mere ownership: The name-letter effect in twelve European languages*, *Eur. J. Soc. Psychol.* 15 (1987), pp. 381–402.
- [6] B.W. Pelham, M.C. Mirenberg, and J.K. Jones, *Why Susie sells seashells by the seashore: Implicit egotism and major life decisions*, *J. Pers. Soc. Psychol.* 82 (2002), pp. 469–487.

Appendix

In the course of replicating the AR1, AR2, and AR3 results, we found a few errors. First, NS miscalculated the "average year of play." They intended to take the arithmetic mean of all the years in which a player played. For example, a player who played from 1930 to 1940 would have 1935 as her average year of play. However, many players are traded during the season, and so have the same year recorded twice (or more!) in the database. These duplicate years need to be eliminated before calculating the average, which NS did not do. The error does not materially affect their results: their reported/actual p -values on p. 1107 are: 0.12/0.012; 0.023/0.028; 0.045/0.044.

We found that the reported p -values are two-sided, not one-sided as befits an hypothesis of this type. It is also worth mentioning that the R^2 measure for the above three ANCOVA results are: 0.098, 0.10, 0.011.

When NS write that they controlled “for whether players were American or foreign born” they actually controlled for *both* birthplace *and* average year of play. Similarly, when they write that they “controlled for country of origin” with dummy variables, they actually controlled for *both* country of origin *and* average year of play.

Finally, for the country dummies we count 41 countries and one category of “birthplace not reported” for a total for 42 dummies. NS reported 52 dummies, but actually used 41 dummies (40 countries and one category of “birthplace not reported”). For this additional result they used $N = 6398$ observations, whereas we used $N = 6397$. Nonetheless, we replicated their p -value.

We also note that NS provide no motivation for including “average year of play,” “American or foreign-born,” or “country of origin” as variables.