Big Macs and Eigenfactor Scores: Don't Let Correlation Coefficients Fool You

Jevin West
Department of Biology, University of Washington, Seattle, WA.
E-mail: jevinw@u.washington.edu

Theodore Bergstrom
Department of Economics, University of California, Santa Barbara, CA

Carl T. Bergstrom
Department of Biology, University of Washington, Seattle, WA
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

The Eigenfactor™ Metrics provide an alternative way of evaluating scholarly journals based on an iterative ranking procedure analogous to Google's PageRank algorithm. These metrics have recently been adopted by Thomson Reuters and are listed alongside the Impact Factor in the *Journal Citation Reports*. But do these metrics differ sufficiently so as to be a useful addition to the bibliometric toolbox? Davis (2008) has argued that they do not, based on his finding of a 0.95 correlation coefficient between Eigenfactor score and Total Citations for a sample of journals in the field of medicine. This conclusion is mistaken; in this article, we illustrate the basic statistical fallacy to which Davis succumbed. We provide a complete analysis of the *2006 Journal Citation Reports* and demonstrate that there are statistically and economically significant differences between the information provided by the Eigenfactor Metrics and that provided by Impact Factor and Total Citations.

Big Macs and Correlation Coefficients

One might think that if the correlation coefficient between two variables is high, those variables convey the same information, and thus can be used interchangeably—but this line of reasoning is erroneous. A simple example helps to illustrate. In Table 1, we provide two statistics for each of 22 countries: the cost of a Big Mac in local currency and the mean hourly wage in local currency. The Pearson product–moment correlation coefficient, ρ, between these two statistics is 0.99. As ρ is nearly 1, one might conclude that we can use hourly wages to predict burger prices with high accuracy and one might question why anyone should waste his or her time collecting burger price information if the hourly wage rates are already known. But take a look at the column "real wage." The real wage—the ratio of burger prices to hourly wages—is the variable of economic interest, as it measures a worker's purchasing power. We see that real wages differ dramatically across countries. In Denmark, a worker making the mean hourly wage need only work for 7 min to earn a Big Mac, whereas in China, a worker making the mean hourly wage must work for nearly 2 h to afford a burger.

In our hamburger example, it is clear what is going on. The denominations of currencies vary immensely and arbitrarily. It is indeed true that differences in real wages are small relative to differences in currency denominations. But it is not true that after correcting for differences in denominations, differences in real wages are negligible. One way to think of this is that the greatest part of the variation in hourly wage comes from the relatively unimportant fact that currency is denominated differently in different countries. The standard deviation of hourly wages in nominal terms is about 300 times as large as that in real terms. Although the standard deviation of real wages across countries is minimal compared with that of nominal exchange rates, this variation is far more important for the quality of life of workers. Thus, one would be wrong to conclude from the high correlation coefficient that the real wage is constant across countries. On the contrary, the standard deviation of this ratio is 62% of the mean.

Davis's Analysis

Davis (2008) fell into a similar trap in his recent comparison of journal rankings by Eigenfactor score and by Impact Factor (IF) or Total Citations. In that study, Davis

Received November 14, 2009; revised April 23, 2010; accepted April 27, 2010

 • Published online 28 May 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.21374

TABLE 1. Hourly Wage versus Real Wage.

Country	Burger Price	Hourly Wage	Real Wage
Denmark	24.75	211.13	8.53
Australia	3.00	19.86	6.62
New Zealand	3.60	21.94	6.09
Switzerland	6.30	37.85	6.01
United States	2.54	14.32	5.64
Britain/UK	1.99	11.15	5.60
Germany	2.61	14.32	5.49
Canada	3.33	16.78	5.04
Singapore	3.30	15.65	4.74
Sweden	24.00	110.90	4.62
Hong Kong	10.70	44.26	4.14
Spain	2.37	8.59	3.62
South Africa	9.70	30.86	3.18
France	2.82	8.50	3.01
Poland	5.90	11.80	2.00
Hungary	399.00	704.34	1.77
Czech Rep.	56.00	85.34	1.52
Brazil	3.60	4.58	1.27
South Korea	3000.00	3134.00	1.04
Mexico	21.90	17.61	0.80
Thailand	55.00	31.69	0.58
China	9.90	5.56	0.56
Mean	166.01	207.32	3.72
Std. dev.	638.49	670.63	2.29
Std. dev./mean	3.85	3.23	0.62

Burger price and hourly wage are in the local currency. Burger price is the average cost of a Big Mac. The units for Real Wage are burgers per hour. Data comes from Behar (2003). The correlation coefficient between burger price and hourly wage is $\rho = 0.99$.

aimed to determine whether measures of "popularity," such as IF and Total Citation, differ substantially from measures of "prestige," such as the journal *PageRank* (Bollen, Rodriquez, & Van de Sompel, 2006) and the Eigenfactor™ Metrics (Bergstrom, 2007).[1] To do so, Davis conducted a regression analysis of Eigenfactor scores on Total Citations[2] for a set of 165 medical journals.[3] Davis reports that the correlation coefficient between 2006 Eigenfactor scores and Total Citations[4] is $\rho = 0.9493$. Based on this result, Davis concluded that

> At least for medical journals, it does not appear that iterative weighting of journals based on citation counts results in rankings that are significantly different from raw citation counts. Or, stated another way, the concepts of popularity (as measured by total citation counts) and prestige (as measured by a weighting mechanism) appear to provide very similar information.

But is Davis right? Is it really the case that if you know the number of citations, you would be wasting your time by finding the Eigenfactor score? Not at all.

First, Davis made a classic statistical error—cautioned against by Karl Pearson (1897)—in comparing two measures with a common factor. Second, Davis suggests that a high correlation coefficient implies that there is no significant difference between two alternative measures; this is simply not true. We address these issues in turn.

Journal Sizes and Spurious Correlations

There are enormous differences in the size of academic journals, and these differences swamp the patterns that Davis was seeking in his analysis. The *Journal Citation Reports* (JCR) indexes journals that range in size from small (*Astronomy and Astrophysics Review* has published 13 articles over the previous 5 years) to big (*The Journal of Biological Chemistry* has published 31,045 articles over the same period) with a coefficient of variation, c_v, equal to 1.910. Per-article citation intensity varies less, whether measured by AI (range 0–27.5, coefficient of variation = 1.785) or by IF (range 0–63.3, coefficient of variation = 1.548).

We can formalize these observations by decomposing Davis's regression of Eigenfactor on Total Citations. Davis regresses

$$\log(\mathrm{EF}_i) \quad \text{versus} \quad \log(\mathrm{CT}_i),$$

where EF_i is the Eigenfactor score for journal i and CT_i the Total Citations received by journal i. We let AI_i be the AI for journal i, and $N_{i,5}$ be the total number of articles published over the last 5 years for journal i. Then, by definition

$$\begin{aligned}\log(\mathrm{EF}_i) &= \log(c_1 \times \mathrm{AI}_i \times N_{i,5}) \\ &= \log c_1 + \log \mathrm{AI}_i + \log N_{i,5},\end{aligned}$$

where c_1 is a scaling constant that normalizes the AI scores so that the mean article in the JCR has an AI score of 1.00. Similarly, letting IF_i be the IF for journal i,

$$\begin{aligned}\log(\mathrm{CT}_i) &\approx \log(c_2 \times \mathrm{IF}_i \times N_{i,2}) \\ &\approx \log(c_2 c_3 \times \mathrm{IF}_i \times N_{i,5}) \\ &= \log c_2 c_3 + \log \mathrm{IF}_i + \log N_{i,5},\end{aligned}$$

[1] The same issue was the subject of a more comprehensive analysis by Bollen et al. (2006). In that study, Bollen et al. compare weighted PageRank with IF and with Total Citations to explore differences between popularity and prestige. Weighted PageRank and Eigenfactor are both variants of the PageRank algorithm. See also Pinski and Narin (1976) for an early attempt at constructing prestige-based measures using citation data, and Vigna (2009) for a discussion of how Pinski and Narin's measure differs from current approaches.

[2] In his study, Davis also looked at the correlation coefficient between Eigenfactor and IF scores. This ρ-value is lower ($\rho = 0.86$), but the point is not so much what this value is, but rather that the comparison makes little sense. Eigenfactor is a measure of total citation impact, and should (all else equal) scale with the size of the journal. IF is a measure of citation impact per article, and all else equal should be independent of journal size. If one wants to compare an Eigenfactor Metric with the IF, one should use the Article Influence (AI) score, which is a per-article measure like IF. We explore this comparison later in the article.

[3] Contrary to what is specified in that study, Davis appears to have sampled from both the "Medicine General and Internal" and "Medicine Research and Experimental" fields, not merely the former category. In our analysis of the same subfields of medicine, we included 168 journals (of the 171 journals in this field); we eliminated 3 journals because they had an IF and/or AI score of zero.

[4] Davis appears to have used citations (from year 2006) to all articles published in the journals he selected. A cleaner comparison, which would have resulted in a higher correlation, would have been to extract citations (from year 2006) to articles published in the past 5 years, as the Eigenfactor score takes into account only the past 5 years' citations.

DOI: 10.1002/asi

where c_2 and c_3 are the additional scaling constants. The scaling constant, c_2, accounts for the fact that Davis compared citations for *all* years and not just citations for 2 years. The scaling constant c_3 relates the number of articles published in 2 years to the number of articles published in 5 years (and thus is approximately 5/2). As a result, Davis is effectively calculating a regression between

$$\log(\text{AI}) + \log(\text{Total Articles})$$

and

$$\log(\text{IF}) + \log(\text{Total Articles}).$$

Having the "log(Total Articles)" term on both the sides of the regression—especially given that it varies more than the other two terms—obscures the relationship between the variables that one would actually wish to observe when trying to evaluate the difference between "popularity" and "prestige."

This pitfall is famous in the history in mathematical statistics. In 1897, two years after pioneering statistician Karl Pearson developed the product–moment correlation coefficient, Pearson presented an article to the Royal Society in which he noted that fellow biometrician W. F. R. Weldon had made precisely this mistake in the analysis of body dimensions of crustaceans (Pearson, 1897; Weldon, 1892). Explaining this error, Pearson wrote

> If the ratio of two absolute measurements on the same or different organs be taken it is convenient to term this ratio an index. If $u = f_1(x, y)$ and $v = f_2(z, y)$ be two functions of the three variables x, y, z, and these variables be selected at random so that there exists no correlation between x,y, y,z, or z,x, there will still be found to exist correlation between u and v. Thus, a real danger arises when a statistical biologist attributes the correlation between two functions, like u and v to organic relationship.

It was to describe this danger that Pearson coined the term *spurious correlation* (Aldrich, 1995; Kronmal, 1993; Pearson, 1897). He imagined a set of bones assembled at random. Based on correlations between measurements that share a common factor, a biologist could easily make the mistake of concluding that the bones were properly assembled into their original skeletons:

> For example, a quantity of bones are taken from an ossuarium, and are put together in groups, which are asserted to be those of individual skeletons. To test this a biologist takes the triplet femur, tibia, humerus, and seeks the correlation between the indices femur/humerus and tibia/humerus. He might reasonably conclude that this correlation marked organic relationship, and believe that the bones had really been put together substantially in their individual grouping. As a matter of fact, since the coefficients of variation for femur, tibia, and humerus are approximately equal, there would be, as we shall see later, a correlation of about 0.4–0.5 between these indices had the bones been sorted absolutely at random. I term this a spurious organic correlation or simply a spurious correlation. I understand by this phrase the amount of correlation that would still exist between the indices, were the absolute lengths on which they depend distributed at random.

The reason for this correlation will be that some of the random femur and tibia pairs will be combined with a large humerus; in this case, both the femur/humerus and tibia/humerus ratio will tend to be smaller than average. Other femur and tibia pairs will be combined with a small humerus; in this case, both the femur/humerus and tibia/humerus ratio will tend to be larger than average. Correlation coefficients of the two ratios give the illusion that tibia and femur length covary, even when they in fact do not. For his part, Weldon was forced to concede that nearly 50% of the correlation he had observed in body measurements was actually due to this effect.

Just over a decade later, another important person in the development of mathematical statics, G. U. Yule, noted that when absolute values share a common factor, they are just as susceptible to this problem as are "indices" or ratios (Yule, 1910):

> Suppose we combine at random two indices z_1 and z_2, e.g. two death rates, and also combine at random with each pair a denominator or population x_3. The correlations between z_1, z_2, and x_3 will then be zero within the limits of sampling. But now suppose we work out the total deaths $x_1 = z_1 x_3$ and $x_2 = z_2 x_3$; the correlation r_{12} between x_1 and x_2 will not be zero, but positive.

This is precisely the form of spurious correlation that arises in Davis's analysis. Per-article popularity as measured by IF takes the role of z_1 in Yule's example, and per-article prestige as measured by AI score takes the role of z_2. Total Articles take the role of Yule's x_3. Even if IF and AI were entirely uncorrelated, Davis still would have observed a high correlation coefficient in his regression of Eigenfactor and Total Citations (approximately $\rho = 0.6$ for all journals), because both share number of articles as a common factor. What Davis discovered is not that popularity and prestige are the same thing; he discovered that big journals are big and small journals are small. Because of this wide variation in journal size, one would also observe a high correlation coefficient between pages and total cites, although very few would argue that the former is an adequate surrogate for the latter.[5]

To avoid this problem, we might want to look at the correlation between popularity *per article* and prestige *per article*. That is, we need to look at the comparison

$$\log(\text{AF}) \quad \text{versus} \quad \log(\text{IF}).$$

Since its inception in January 2007, Eigenfactor.org has provided exactly this information at http://www.eigenfactor.org/correlation/, for the entire JCR dataset and also for each individual field of scholarship as defined by the JCR.[6]

[5]We collected page and citation information for 149 Economics journals in 2006. The correlation coefficient between total pages and total citations is $\rho = 0.615$.

[6]Falagas, Kouranos, Arencibia-Jorge, and Karageorgopoulos (2008) presented a similar comparison of IF and the Scimago journal rank indicator (a per-article measure of prestige). Waltman and van Eck (2010) look at correlations among a number of bibliometric measures; their discussion of differences between IF and AI is noteworthy.

DOI: 10.1002/asi

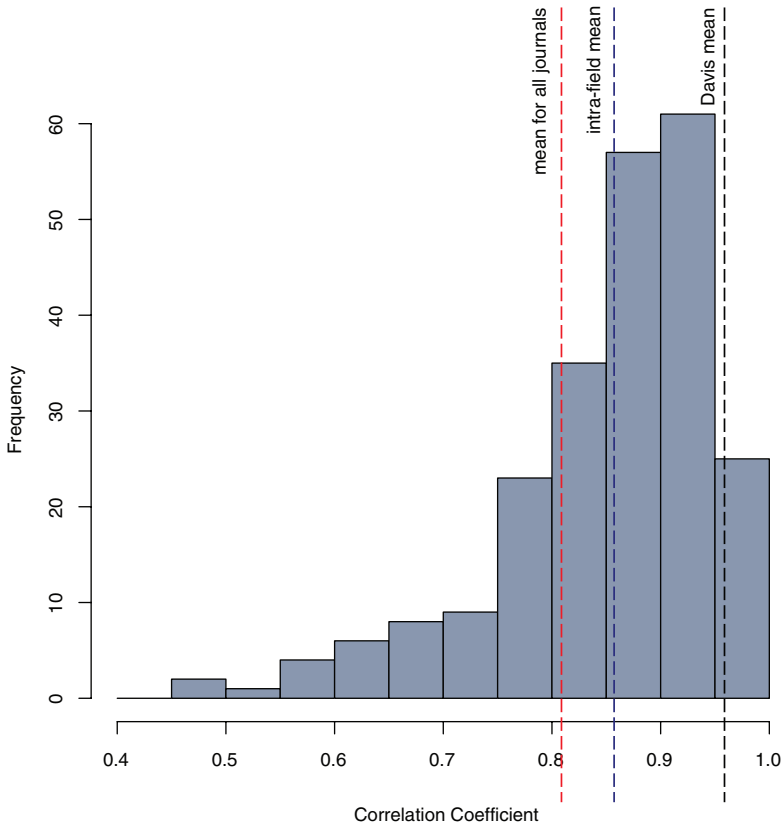

FIG. 1. Histogram of correlation coefficients between IF and AI scores. This includes all 231 categories in the 2006 Science and Social Science JCR. The mean of all fields is 0.853 (intra-field mean) and the standard deviation is 0.099. The correlation for all journals considered together is 0.818. The correlation for the field of Medicine as studied by Davis is 0.954. The correlation coefficients for all fields can be found at http:/www.eigenfactor.org/correlation/

Figure 1 is a histogram of the correlation coefficients between IF and AI scores for all 231 categories in the 2006 JCR. The mean for all the fields was 0.853 with a standard deviation of 0.099. The field with the lowest correlation coefficient is Communication ($\rho = 0.478$). Marine Engineering has the highest correlation ($\rho = 0.986$). The sample of medical journals that Davis selected, with $\rho = 0.954$, ranks in the 90th percentile when compared with all 231 fields. Correlation coefficients within fields typically exceed the correlation coefficient for all the journals together. For all 7,611 journals considered together, $\rho = 0.818$. This value is lower than the mean of individual-field correlation coefficients, which is $\rho = 0.853$.

Correlation and Significant Differences

To evaluate Davis's claim that Eigenfactor score and Total Citations are telling us the same thing, we can focus on the *ratio* of Eigenfactor score to Total Citations (EF/TC). (When we look at the ratio, the common factor "Total Articles" divides out.) Notice that a journal's EF/TC ratio is a measure of "bang per cite received"—that is, how much Eigenfactor boost does this journal receive, on average, when it is cited. In the hamburger example, the corresponding notion is "burgers per hour," the real wage or purchasing power of an hour's work. Does a high correlation between Total Citations and Eigenfactor score mean that the bang per cite received is about constant? If it is, there really would be no point in looking at Eigenfactor scores instead of Total Citations. Hence, let's see what happens.

Figure 2 shows the ratio of Eigenfactor score to Total Citations for every journal in the JCR, and the inset shows just the medical journals. The standard deviation of this ratio is 1.1×10^{-5} and the mean is 1.56×10^{-5}. The standard deviation, in this case, is 71% of the mean. This is even more variable than the Big Mac case. Moreover, there are nearly 1,000 journals with twice the mean "bang per cite."

The thing to notice in both the Big Mac and the journal example is that if you are interested in the ratio of A to B and if $A = ax$ and $B = bx$ for some x with a very high variance relative to that of a and of b, you will get a very high ρ value when you regress B on A. However, if what really interests you is the ratio A/B, you will note that the x's cancel and $A/B = ax/bx = a/b$. Thus, the variance of x has literally nothing to state about the variance of the ratio a/b. You do not learn about whether a/b is nearly constant or highly variable from looking at the correlation of B on A.

If, as Davis claims, Eigenfactor scores do not differ significantly from Total Citation counts, the ratio EF/TC should be constant across different groups of journals. To evaluate

DOI: 10.1002/asi

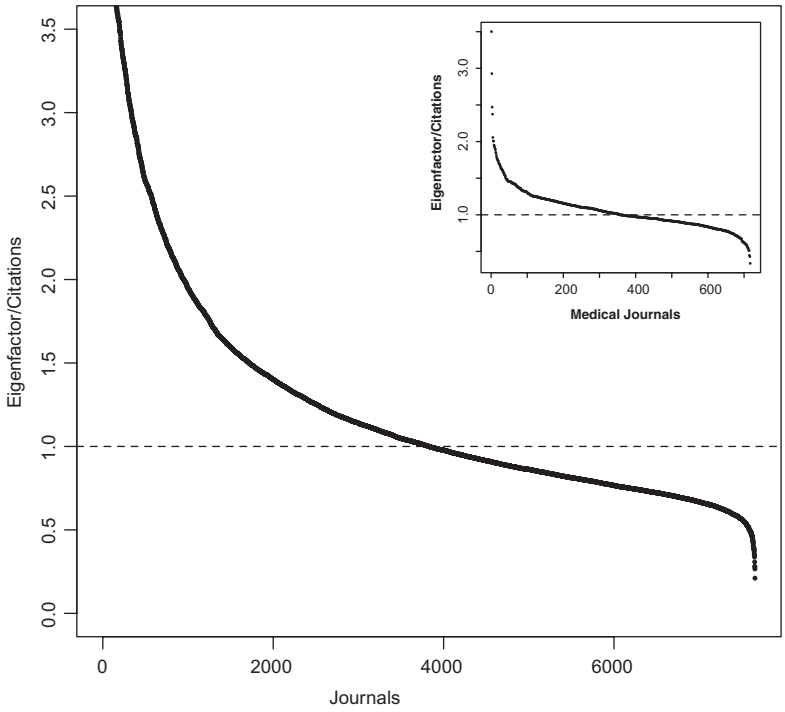

FIG. 2. Ratio of Eigenfactor score to Total Citations. Data are normalized by the median ratio of the dataset. The dashed line indicates a ratio of 1. The journals are ordered from those with the highest ratio to the lowest. The inset shows only the 168 medical journals from Davis's analysis.

this claim, we look at the EF/TC ratios of social journals with those of science journals, with groupings determined by whether a journal is listed in the Social Science JCR or the Science JCR. (Journals listed in both are omitted from the analysis.) The mean EF/TC ratio for science journals is 1.42×10^{-5}, whereas the mean for social science journals is 2.12×10^{-5}. A Mann–Whitney U-test shows that this difference is highly significant, at the $p < 10^{-167}$ level.

These differences are not only statistically significant, but also economically relevant. The 49% difference in mean EF/TC ratios indicates that a librarian who uses Total Citations to measure journal value will underestimate the value of social science journals by 49% relative to a librarian who uses Eigenfactor scores to measure value.

There are also significant differences within the sample of journals that Davis considered. Based on the difference between science and social science ratios described earlier, one might expect medical journals more closely associated with the social sciences, such as those in public health, to have higher-than-average EF/TC ratios. Seven of the publications in Davis's sample of medical journals are cross-listed in the JCR category of public, environmental, and occupational health. Indeed, this group of journals has a 29% higher EF/TC ratio than do the rest of the journals in Davis's sample, again statistically significant (Mann–Whitney U-test, $p < 0.01$).

Note that there is nothing special about this particular comparison between sciences and social sciences; one could test any number of alternative hypotheses and would find significant differences between EF/TC ratios for many other comparisons as well.

The Value of Visualization

Hence, if correlation coefficients are misleading, what is the alternative? First, we argue for a deeper examination of the data. Figure 3 is an example of this strategy. Listing the journals in this way, one can quickly see the ordinal differences that exist between these highly correlated data. This type of graphical display illustrates the interesting stories that can be lost behind a summary statistic, such as the Spearman correlation.

Figure 3 illustrates the ordinal ranks of the top 50% of the medical journals used in Davis's study. In the left column, the journals in this subfield of medicine are ranked by the total number of citations. In the right column, the journals are ordered by the Eigenfactor score. The lines connecting the journals indicate whether the journal moved up (green), down (red), or stayed the same (black) relative to their ranking by Total Citations. The figure highlights the differences between the metrics. For example, *Aviation Space and Environmental Medicine* drops 30 places, whereas *PLoS Medicine* raises 31 places. Davis claims in his study that the ordering of journals does not change drastically. Figure 3 suggests otherwise.

Figure 4 compares the ordinal ranking by IF and AI for 84 journals—the top-ranked half—from Davis's study. Changes in ranking are even more dramatic when we look at the lower-ranked 84 journals. The correlation coefficient between IF and AI for the top 84 journals is $\rho = 0.955$. Despite this high

DOI: 10.1002/asi

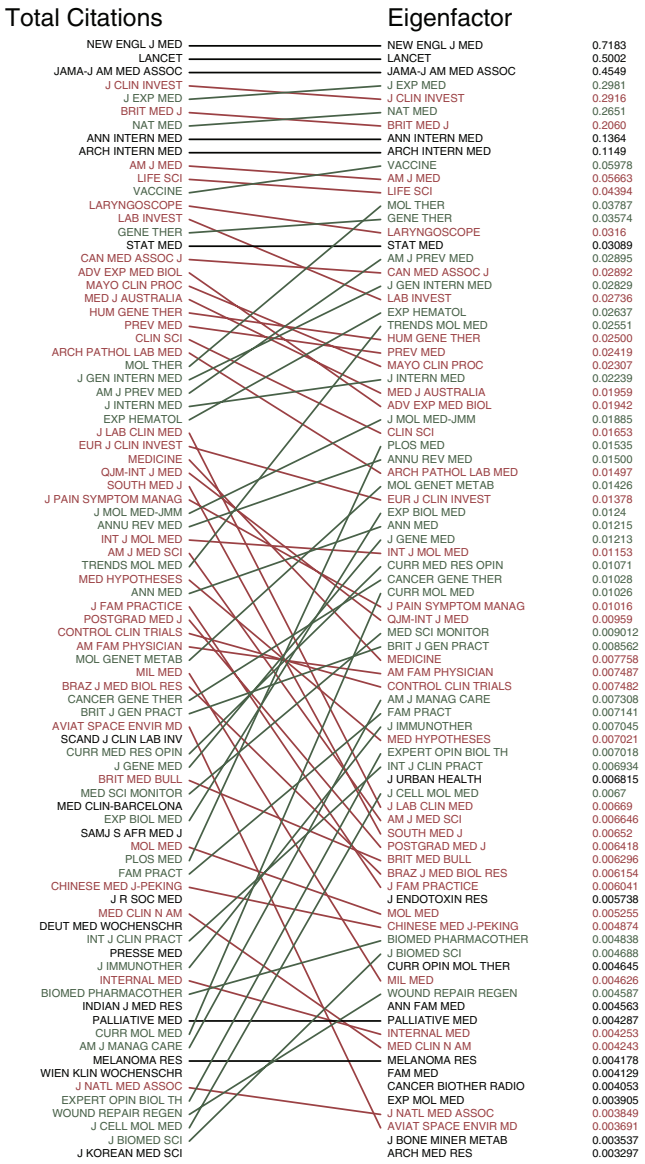

FIG. 3. Journal ranking comparisons by Total Citations and Eigenfactor score. The journals listed are the top 50% from the field of Medicine that Davis analyzed. Journals in the left column are ranked by Total Citations for all years. Journals in the right column are ranked by Eigenfactor score. The lines connecting the journals indicate whether the journal moved up (green), down (red), or stayed the same (black) relative to their ranking by Total Citations. Journal names in black can also be journals that do not exist in both the columns.

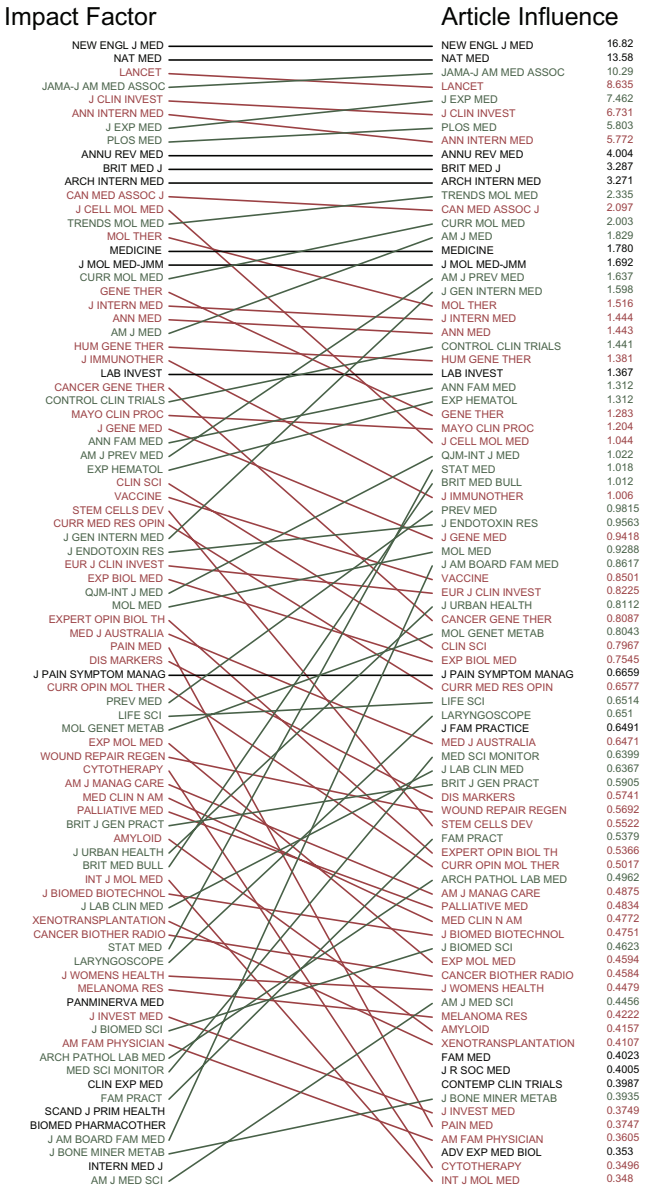

FIG. 4. Comparing IF and AI. The journals shown are from the same field that Davis analyzed (because of limited space, only the top 84 journals are shown). For these 84 journals, the correlation coefficient between IF and AI is $\rho = 0.955$. The relative rankings by IF and AI are listed in the left and right column, respectively. The third column lists the AI scores. The journal names in green indicate those that fare better when ranked by AI; the journal names in red fare better when ranked by IF. The names in black are journals that exhibit no change or exist outside the range of the journals shown.

correlation, the figure highlights the fact that the two metrics yield substantially different ordinal rankings.

Figure 4 reveals that the top few journals change in rank less than those further down the hierarchy. For example, going from IF to AI, the journals in the top 10 change in rank by only 1 or 2 positions. By contrast, there are many larger changes further on in the rankings.[7] For example, as we go from IF to AI, the *Journal of General Internal Medicine* rises 18 spots to number 19, whereas *Pain Medicine* drops 35 spots to end up at number 80. These are just two of the many major shifts (in a field with a correlation of 0.955!). These changes in relative ranking would certainly not go unnoticed by editors or publishers.

Furthermore, while ordinal changes are interesting, cardinal changes are often more important. Figure 5 shows the top 10 journals from Figure 4—those with the least ordinal change from one metric to another—now in their

[7]Bollen et al. (2006) observed a similar pattern in a series of scatterplots contrasting PageRank and IF values for all journals. In these scatterplots, the rankings of top-tier journals differ relatively little, whereas more variation is found in the middle and bottom portions of the hierarchy.

DOI: 10.1002/asi

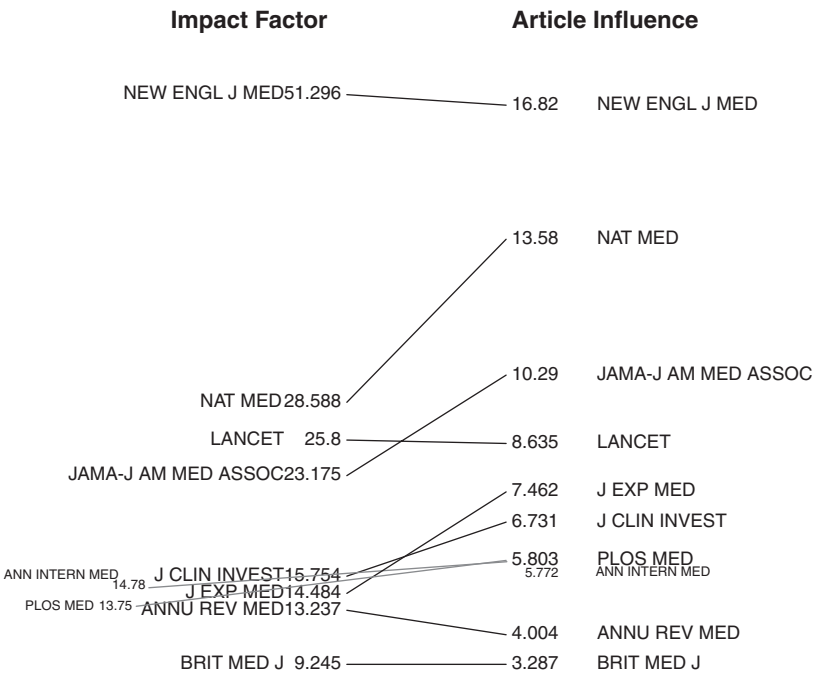

FIG. 5. Cardinal differences between IF and AI scores. The top 10 journals by IF are shown in the left column. The scores are scaled vertically, reflecting their cardinal positions. The smallest IF score is on the bottom, and the highest IF score is on the top. The right column shows the same journals scaled by AI.

cardinal positions. Even those journals that do not change ordinal rank from one metric to another may be valued very differently under the two different metrics. For example, *Nature Medicine* is the #2 journal regardless of whether one uses IF or AI. But under IF, it has barely half the prestige of the first-place *New England Journal of Medicine*, whereas by AI it makes up a good deal of that ground.

Conclusion

Correlation coefficients can be useful statistical tools. They can help us identify some kinds of statistically significant relationships between pairs of variables, and they can tell us about the sign (positive or negative) of these relationships. One must use considerably greater caution, however, when drawing conclusions from the magnitude of correlation coefficients—all the more so in the presence of spurious correlates and in the absence of a formal hypothesis-testing framework. In particular, we have illustrated that just because two metrics have a high correlation—0.8 or 0.9 or even higher—we cannot safely conclude that they convey the same information, or that one has little additional information to tell us beyond what we learn from the other.

Comparative studies of alternative measures can be very useful in choosing an appropriate bibliometric toolkit. We close with a few suggestions for how one might better conduct these sorts of analyses. First, be wary of what correlation coefficients say about the relationship of two metrics (Anscombe, 1973; Tukey, 1954). High correlation does *not* necessarily mean that two variables provide the same information any more than a low correlation means that two variables are unrelated. Purchasing power varies wildly despite the high correlation between wage and hamburger price in our Big Mac example. At the other end of the spectrum, in the chaotic region of the logistic map, successive iterates have an immediate algebraic relationship yet a correlation of zero.

Second, appropriate data visualization can bring out facets of the data that are obscured by summary statistics. Different forms of data graphics can be better suited for certain tasks; for example, the comparison plots, such as those in Figure 4, better highlight the differences between bibliometric measures than do standard scatter plots.

Finally, simple observations can be at least as powerful as rote statistical calculations in understanding the nature of our data. For example, the median of the burgers per hour in the top third of the countries is about five times the median of the burgers per hour in the bottom third. This explains a great deal about the differences in purchasing power across countries. The median "bang per cite received" in the top third of journals is almost 2.4 times of the median in the bottom third. This says a great deal about the difference in how journals are valued under the Eigenfactor Metrics, and helps us understand why the Eigenfactor Metrics offer a substantially different view of journal prestige than that which we get from straight citation counts.

Acknowledgments

The authors thank Ben Althouse for assistance with Figures 3, 5, and 6, Cosma Shalizi for the helpful discussions, Johan Bollen for the extensive feedback on the manuscript, and an anonymous reviewer for provocative commentary. This research was supported in part by NSF grant SBE-0915005 to C. T. B.

DOI: 10.1002/asi

References

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. Statistical Science, 10(4), 364–376.

Anscombe, F.J. (1973). Graphs in statistical analysis. American Statistician, 27(1), 17–21.

Behar, A. (2003). Who earns the most hamburgers per hour? Retrieved May 6, 2010, from http://www.economics.ox.ac.uk/members/alberto.behar/rw/Burgers.pdf

Bergstrom, C.T. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. College and Research Libraries News, 68(5), 314–316.

Bollen, J., Rodriquez, M.A., & Van de Sompel, H. (2006). Journal status. Scientometrics, 69(3), 669–687.

Davis, P.M. (2008). Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? Journal of the American Society for Information Science and Technology, 59(13), 2186–2188.

Falagas, M.E., Kouranos, V.D., Arencibia-Jorge, R., & Karageorgopoulos, D.E. (2008). Comparison of Scimago journal rank indicator with journal impact factor. The FASEB Journal, 22(8), 2623–2628.

Kronmal, R.A. (1993). Spurious correlation and the fallacy of the ratio standard revisited. Journal of the Royal Statistical Society. Series A (Statistics in Society), 156(3), 379–392.

Pearson, K. (1897). Mathematical contributions to the theory of evolution—On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London, 60, 489–498.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. Information Processing and Management, 12(5), 297–326.

Tukey, J.W. (1954). Unsolved problems of experimental statistics. Journal of the American Statistical Association, 49(268), 706–731.

Vigna, S. (2009). Spectral ranking. Retrieved May 5, 2010, from http://vigna.dsi.unimi.it/papers.php

Waltman, L., & van Eck, N.J. (2010). The relation between eigenfactor, audience factor, and influence weight. Retrieved May 5, 2010, from http://arxiv.org/abs/1003.2198v1

Weldon, F.R.S. (1892). Certain correlated variations in *Crangon vulgaris*. Proceedings of the Royal Society of London, 51, 1–21.

Yule, G.U. (1910). On the interpretation of correlations between indices or ratios. Journal of the Royal Statistical Society, 63(6/7), 644–647.

DOI: 10.1002/asi