An index to quantify an individual's scientific research output

J. E. Hirsch*

Department of Physics, University of California at San Diego, La Jolla, CA 92093-0319

Communicated by Manuel Cardona, Max Planck Institute for Solid State Research, Stuttgart, Germany, September 1, 2005 (received for review August 15, 2005)

I propose the index *h*, defined as the number of papers with citation number ≥*h*, as a useful index to characterize the scientific output of a researcher.

citations | impact | unbiased

For the few scientists who earn a Nobel prize, the impact and relevance of their research is unquestionable. Among the rest of us, how does one quantify the cumulative impact and relevance of an individual's scientific research output? In a world of limited resources, such quantification (even if potentially distasteful) is often needed for evaluation and comparison purposes (e.g., for university faculty recruitment and advancement, award of grants, etc.).

The publication record of an individual and the citation record clearly are data that contain useful information. That information includes the number (N_p) of papers published over n years, the number of citations (N_c^j) for each paper (j), the journals where the papers were published, their impact parameter, etc. This large amount of information will be evaluated with different criteria by different people. Here, I would like to propose a single number, the "h index," as a particularly simple and useful way to characterize the scientific output of a researcher.

A scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each.

The research reported here concentrated on physicists; however, I suggest that the h index should be useful for other scientific disciplines as well. (At the end of the paper I discuss some observations for the h index in biological sciences.) The highest h among physicists appears to be E. Witten's h, which is 110. That is, Witten has written 110 papers with at least 110 citations each. That gives a lower bound on the total number of citations to Witten's papers at $h^2 = 12{,}100$. Of course, the total number of citations ($N_{c,tot}$) will usually be much larger than h^2, because h^2 both underestimates the total number of citations of the h most-cited papers and ignores the papers with $<h$ citations. The relation between $N_{c,tot}$ and h will depend on the detailed form of the particular distribution (1), and it is useful to define the proportionality constant a as

$$N_{c,tot} = ah^2. \qquad [1]$$

I find empirically that a ranges between 3 and 5.

Other prominent physicists with high hs are A. J. Heeger ($h = 107$), M. L. Cohen ($h = 94$), A. C. Gossard ($h = 94$), P. W. Anderson ($h = 91$), S. Weinberg ($h = 88$), M. E. Fisher ($h = 88$), M. Cardona ($h = 86$), P. G. deGennes ($h = 79$), J. N. Bahcall ($h = 77$), Z. Fisk ($h = 75$), D. J. Scalapino ($h = 75$), G. Parisi ($h = 73$), S. G. Louie ($h = 70$), R. Jackiw ($h = 69$), F. Wilczek ($h = 68$), C. Vafa ($h = 66$), M. B. Maple ($h = 66$), D. J. Gross ($h = 66$), M. S. Dresselhaus ($h = 62$), and S. W. Hawking ($h = 62$). I argue that h is preferable to other single-number criteria commonly used to evaluate scientific output of a researcher, as follows:

(i) Total number of papers (N_p). Advantage: measures productivity. Disadvantage: does not measure importance or impact of papers.

(ii) Total number of citations ($N_{c,tot}$). Advantage: measures total impact. Disadvantage: hard to find and may be inflated by a small number of "big hits," which may not be representative of the individual if he or she is a coauthor with many others on those papers. In such cases, the relation in Eq. **1** will imply a very atypical value of a, >5. Another disadvantage is that $N_{c,tot}$ gives undue weight to highly cited review articles versus original research contributions.

(iii) Citations per paper (i.e., ratio of $N_{c,tot}$ to N_p). Advantage: allows comparison of scientists of different ages. Disadvantage: hard to find, rewards low productivity, and penalizes high productivity.

(iv) Number of "significant papers," defined as the number of papers with $>y$ citations (for example, $y = 50$). Advantage: eliminates the disadvantages of criteria i, ii, and iii and gives an idea of broad and sustained impact. Disadvantage: y is arbitrary and will randomly favor or disfavor individuals, and y needs to be adjusted for different levels of seniority.

(v) Number of citations to each of the q most-cited papers (for example, $q = 5$). Advantage: overcomes many of the disadvantages of the criteria above. Disadvantage: It is not a single number, making it more difficult to obtain and compare. Also, q is arbitrary and will randomly favor and disfavor individuals.

Instead, the proposed h index measures the broad impact of an individual's work, avoids all of the disadvantages of the criteria listed above, usually can be found very easily by ordering papers by "times cited" in the Thomson ISI Web of Science database (http://isiknowledge.com),† and gives a ballpark estimate of the total number of citations (Eq. **1**).

Thus, I argue that two individuals with similar hs are comparable in terms of their overall scientific impact, even if their total number of papers or their total number of citations is very different. Conversely, comparing two individuals (of the same scientific age) with a similar number of total papers or of total citation count and very different h values, the one with the higher h is likely to be the more accomplished scientist.

For a given individual, one expects that h should increase approximately linearly with time. In the simplest possible model, assume that the researcher publishes p papers per year and that each published paper earns c new citations per year every subsequent year. The total number of citations after $n + 1$ years is then

$$N_{c,tot} = \sum_{j=1}^{n} pcj = \frac{pcn(n+1)}{2}. \qquad [2]$$

*E-mail: jhirsch@ucsd.edu.

†Of course, the database used must be complete enough to cover the full period spanned by the individual's publications.

© 2005 by The National Academy of Sciences of the USA

Downloaded at GRANADA/F CIENCIAS on August 25, 2020

PHYSICS

Assuming all papers up to year y contribute to the index h, we have

$$(n - y)c = h \quad \textbf{[3a]}$$

$$py = h. \quad \textbf{[3b]}$$

The left side of Eq. **3a** is the number of citations to the most recent of the papers contributing to h; the left side of Eq. **3b** is the total number of papers contributing to h. Hence, from Eq. **3**,

$$h = \frac{c}{1 + c/p} n. \quad \textbf{[4]}$$

The total number of citations (for not-too-small n) is then approximately

$$N_{c,tot} \sim \frac{(1 + c/p)^2}{2c/p} h^2 \quad \textbf{[5]}$$

of the form Eq. **1**. The coefficient a depends on the number of papers and the number of citations per paper earned per year as given by Eq. **5**. As stated earlier, we find empirically that $a \approx 3$–5 is a typical value. The linear relation

$$h \sim mn \quad \textbf{[6]}$$

should hold quite generally for scientists who produce papers of similar quality at a steady rate over the course of their careers; of course, m will vary widely among different researchers. In the simple linear model, m is related to c and p as given by Eq. **4**. Quite generally, the slope of h versus n, the parameter m, should provide a useful yardstick to compare scientists of different seniority.

In the linear model, the minimum value of a in Eq. **1** is $a = 2$, for the case $c = p$, where the papers with $>h$ citations and those with $<h$ citations contribute equally to the total $N_{c,tot}$. The value of a will be larger for both $c > p$ and $c < p$. For $c > p$, most contributions to the total number of citations arise from the "highly cited papers" (the h papers that have $N_c > h$), whereas for $c < p$, it is the sparsely cited papers (the $N_p - h$ papers that have $<h$ citations each) that give the largest contribution to $N_{c,tot}$. We find that the first situation holds in the vast majority of, if not all, cases. For the linear model defined in this example, $a = 4$ corresponds to $c/p = 5.83$ (the other value that yields $a = 4$, $c/p = 0.17$, is unrealistic).

The linear model defined above corresponds to the distribution

$$N_c(y) = N_0 - \left(\frac{N_0}{h} - 1\right) y, \quad \textbf{[7]}$$

where $N_c(y)$ is the number of citations to the yth paper (ordered from most cited to least cited) and N_0 is the number of citations of the most highly cited paper ($N_0 = cn$ in the example above). The total number of papers y_m is given by $N_c(y_m) = 0$; hence,

$$y_m = \frac{N_0 h}{N_0 - h}. \quad \textbf{[8]}$$

We can write N_0 and y_m in terms of a defined in Eq. **1** as

$$N_0 = h\left[a \pm \sqrt{a^2 - 2a}\right] \quad \textbf{[9a]}$$

$$y_m = h\left[a \mp \sqrt{a^2 - 2a}\right]. \quad \textbf{[9b]}$$

For $a = 2$, $N_0 = y_m = 2h$. For larger a, the upper sign in Eq. **9** corresponds to the case where the highly cited papers dominate (the more realistic case), and the lower sign corresponds to the case where the less frequently cited papers dominate the total citation count.

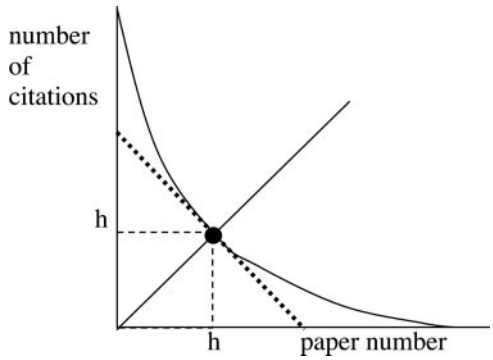

Fig. 1. Schematic curve of number of citations versus paper number, with papers numbered in order of decreasing citations. The intersection of the 45° line with the curve gives h. The total number of citations is the area under the curve. Assuming the second derivative is nonnegative everywhere, the minimum area is given by the distribution indicated by the dotted line, yielding $a = 2$ in Eq. **1**.

In a more realistic model, $N_c(y)$ will not be a linear function of y. Note that $a = 2$ can safely be assumed to be a lower bound quite generally, because a smaller value of a would require the second derivative $\partial^2 N_c/\partial y^2$ to be negative over large regions of y, which is not realistic. The total number of citations is given by the area under the $N_c(y)$ curve that passes through the point $N_c(h) = h$. In the linear model, the lowest $a = 2$ corresponds to the line of slope -1, as shown in Fig. 1.

A more realistic model would be a stretched exponential of the form

$$N_c(y) = N_0 e^{-\left(\frac{y}{y_0}\right)^\beta}. \quad \textbf{[10]}$$

Note that for $\beta \leq 1$, $N_c''(y) > 0$ for all y; hence, $a > 2$ is true. We can write the distribution in terms of h and a as

$$N_c(y) = \frac{a}{\alpha I(\beta)} h e^{-\left(\frac{y}{h\alpha}\right)^\beta} \quad \textbf{[11]}$$

with $I(\beta)$ the integral

$$I(\beta) = \int_0^\infty dz e^{-z^\beta} \quad \textbf{[12]}$$

and α determined by the equation

$$\alpha e^{\alpha - \beta} = \frac{a}{I(\beta)}. \quad \textbf{[13]}$$

The maximally cited paper has citations

$$N_0 = \frac{a}{\alpha I(\beta)} h, \quad \textbf{[14]}$$

and the total number of papers (with at least one citation) is determined by $N(y_m) = 1$ as

$$y_m = h[1 + \alpha^\beta \ln(h)]^{1/\beta}. \quad \textbf{[15]}$$

A given researcher's distribution can be modeled by choosing the most appropriate β and a for that case. For example, for $\beta = 1$, if $a = 3$, $\alpha = 0.661$, $N_0 = 4.54h$, and $y_m = h[1 + .66\ln h]$. With $a = 4$, $\alpha = 0.4644$, $N_0 = 8.61h$, and $y_m = h[1 + 0.46\ln(h)]$. For

Downloaded at GRANADA/F CIENCIAS on August 25, 2020

$\beta = 0.5$, the lowest possible value of a is 3.70; for that case, $N_0 = 7.4h$ and $y_m = h[1 + 0.5\ln(h)]^2$. Larger a values will increase N_0 and reduce y_m. For $\beta = 2/3$, the smallest possible a is $a = 3.24$, for which case $N_0 = 4.5h$ and $y_m = h[1 + 0.66\ln(h)]^{3/2}$.

The linear relation between h and n (Eq. **6**) will of course break down when the researcher slows down in paper production or stops publishing altogether. There is a time lag between the two events. In the linear model, assuming the researcher stops publishing after $n_{\rm stop}$ years, h continues to increase at the same rate for a time

$$n_{lag} = \frac{h}{c} = \frac{1}{1 + c/p}\, n_{\rm stop} \qquad [16]$$

and then stays constant, because now all published papers contribute to h. In a more realistic model, h will smoothly level off as n increases rather than with a discontinuous change in slope. Still, quite generally, the time lag will be larger for scientists who have published for many years, as Eq. **16** indicates.

Furthermore, in reality, of course, not all papers will eventually contribute to h. Some papers with low citations will never contribute to a researcher's h, especially if written late in the career, when h is already appreciable. As discussed by Redner (3), most papers earn their citations over a limited period of popularity and then they are no longer cited. Hence, it will be the case that papers that contributed to a researcher's h early in his or her career will no longer contribute to h later in the individual's career. Nevertheless, it is of course always true that h cannot decrease with time. The paper or papers that at any given time have exactly h citations are at risk of being eliminated from the individual's h count as they are superseded by other papers that are being cited at a higher rate. It is also possible that papers "drop out" and then later come back into the h count, as would occur for the kind of papers termed "sleeping beauties" (4).

For the individual researchers mentioned earlier, I find n from the time elapsed since their first published paper till the present and find the following values for the slope m defined in Eq. **6**: Witten, $m = 3.89$; Heeger, $m = 2.38$; Cohen, $m = 2.24$; Gossard, $m = 2.09$; Anderson, $m = 1.88$; Weinberg, $m = 1.76$; Fisher, $m = 1.91$; Cardona, $m = 1.87$; deGennes, $m = 1.75$; Bahcall, $m = 1.75$; Fisk, $m = 2.14$; Scalapino, $m = 1.88$; Parisi, $m = 2.15$; Louie, $m = 2.33$; Jackiw, $m = 1.92$; Wilczek, $m = 2.19$; Vafa, $m = 3.30$; Maple, $m = 1.94$; Gross, $m = 1.69$; Dresselhaus, $m = 1.41$; and Hawking, $m = 1.59$. From inspection of the citation records of many physicists, I conclude the following:

(*i*) A value of $m \approx 1$ (i.e., an h index of 20 after 20 years of scientific activity), characterizes a successful scientist.
(*ii*) A value of $m \approx 2$ (i.e., an h index of 40 after 20 years of scientific activity), characterizes outstanding scientists, likely to be found only at the top universities or major research laboratories.
(*iii*) A value of $m \approx 3$ or higher (i.e., an h index of 60 after 20 years, or 90 after 30 years), characterizes truly unique individuals.

The m parameter ceases to be useful if a scientist does not maintain his or her level of productivity, whereas the h parameter remains useful as a measure of cumulative achievement that may continue to increase over time even long after the scientist has stopped publishing.

Based on typical h and m values found, I suggest (with large error bars) that for faculty at major research universities, $h \approx 12$ might be a typical value for advancement to tenure (associate professor) and that $h \approx 18$ might be a typical value for advancement to full professor. Fellowship in the American Physical Society might occur typically for $h \approx 15$–20. Membership in the National Academy of Sciences of the United States of America may typically be associated with $h \approx 45$ and higher, except in exceptional circumstances. Note that these estimates correspond roughly to the typical number of years of sustained research production assuming an $m \approx 1$ value; the time scales of course will be shorter for scientists with higher m values. Note that the time estimates are taken from the publication of the first paper, which typically occurs some years before the Ph.D. is earned.

There are, however, a number of caveats that should be kept in mind. Obviously, a single number can never give more than a rough approximation to an individual's multifaceted profile, and many other factors should be considered in combination in evaluating an individual. Furthermore, the fact that there can always be exceptions to rules should be kept in mind, especially in life-changing decisions such as the granting or denying of tenure. There will be differences in typical h values in different fields, determined in part by the average number of references in a paper in the field, the average number of papers produced by each scientist in the field, and the size (number of scientists) of the field (although, to a first approximation in a larger field, there are more scientists to share a larger number of citations, so typical h values should not necessarily be larger). Scientists working in nonmainstream areas will not achieve the same very high h values as the top echelon of those working in highly topical areas. Although I argue that a high h is a reliable indicator of high accomplishment, the converse is not necessarily always true. There is considerable variation in the skewness of citation distributions even within a given subfield, and for an author with a relatively low h that has a few seminal papers with extraordinarily high citation counts, the h index will not fully reflect that scientist's accomplishments. Conversely, a scientist with a high h achieved mostly through papers with many coauthors would be treated overly kindly by his or her h. Subfields with typically large collaborations (e.g., high-energy experiment) will exhibit larger h values, and I suggest that in cases of large differences in the number of coauthors, it may be useful in comparing different individuals to normalize h by a factor that reflects the average number of coauthors. For determining the scientific "age" in the computation of m, the very first paper may sometimes not be the appropriate starting point if it represents a relatively minor early contribution well before sustained productivity ensued.

Finally, in any measure of citations, ideally one would like to eliminate the self-citations. Although self-citations can obviously increase a scientist's h, their effect on h is much smaller than on the total citation count. First, all self-citations to papers with $<h$ citations are irrelevant, as are the self-citations to papers with many more than h citations. To correct h for self-citations, one would consider the papers with number of citations just $>h$ and count the number of self-citations in each. If a paper with $h + n$ citations has $>n$ self-citations, it would be dropped from the h count, and h would drop by 1. Usually, this procedure would involve very few if any papers. As the other face of this coin, scientists intent in increasing their h index by self-citations would naturally target those papers with citations just $<h$.

As an interesting sample population, I computed h and m for the physicists who obtained Nobel prizes in the last 20 years (for calculating m, I used the latter of the first published paper year or 1955, the first year in the ISI database). However, the set was further restricted by including only the names that uniquely identified the scientist in the ISI citation index, which restricted our set to 76% of the total. It is, however, still an unbiased estimator, because the commonality of the name should be uncorrelated with h and m. h indices range from 22 to 79, and m indices range from 0.47 to 2.19. Averages and standard deviations are $\langle h \rangle = 41$, $\sigma_h = 15$ and $\langle m \rangle = 1.14$, $\sigma_m = 0.47$. The distribution of h indices is shown in Fig. 2; the median is at $h_m =$

Downloaded at GRANADA/F CIENCIAS on August 25, 2020

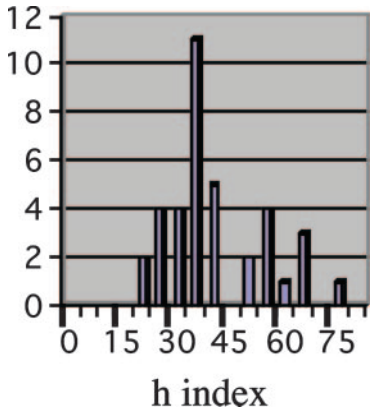

Fig. 2. Histogram giving the number of Nobel prize recipients in physics in the last 20 years versus their *h* index. The peak is at the *h* index between 35 and 39.

35, lower than the mean due to the tail for high *h* values. It is interesting that Nobel prize winners have substantial *h* indices (84% had an *h* of at least 30), indicating that Nobel prizes do not originate in one stroke of luck but in a body of scientific work. Notably, the values of *m* found are often not high compared with other successful scientists (49% of our sample had $m < 1$), clearly because Nobel prizes are often awarded long after the period of maximum productivity of the researchers.

As another example, among newly elected members of the National Academy of Sciences in physics and astronomy in 2005, I find $\langle h \rangle = 44$, $\sigma_h = 14$, highest $h = 71$, lowest $h = 20$, and median $h_m = 46$. Among the total membership in the National Academy of Sciences in physics, the subgroup of last names starting with "A" and "B" has $\langle h \rangle = 38$, $\sigma_h = 10$, and $h_m = 37$. These examples further indicate that the index *h* is a stable and consistent estimator of scientific achievement.

An intriguing idea is the extension of the *h*-index concept to groups of individuals.[‡] The SPIRES high-energy physics literature database (www.slac.stanford.edu/spires/hep) recently implemented the *h* index in their citation summaries, and it also allows the computation of *h* for groups of scientists. The overall *h* index of a group will generally be larger than that of each of the members of the group but smaller than the sum of the individual *h* indices, because some of the papers that contribute to each individual's *h* will no longer contribute to the group's *h*. For example, the overall *h* index of the condensed matter group at the University of California at San Diego physics department is $h = 118$, of which the largest individual contribution is 25; the highest individual *h* is 66, and the sum of individual *h*s is >300. The contribution of each individual to the group's *h* is not necessarily proportional to the individual's *h*, and the highest contributor to the group's *h* will not necessarily be the individual with highest *h*. In fact, in principle (although rarely in practice), the lowest-*h* individual in a group could be the largest contributor to the group's *h*. For a prospective graduate student considering different graduate programs, a ranking of groups or departments in his or her chosen area according to their overall *h* index would likely be of interest, and for administrators concerned with these issues, the ranking of their departments or entire institution according to the overall *h* could also be of interest.

To conclude, I discuss some observations in the fields of biological and biomedical sciences. From the list compiled by Christopher King of Thomson ISI of the most highly cited scientists in the period 1983–2002 (5), I found the *h* indices for the top 10 on that list, all in the life sciences, which are, in order of decreasing *h*: S. H. Snyder, $h = 191$; D. Baltimore, $h = 160$; R. C. Gallo, $h = 154$; P. Chambon, $h = 153$; B. Vogelstein, $h = 151$; S. Moncada, $h = 143$; C. A. Dinarello, $h = 138$; T. Kishimoto, $h = 134$; R. Evans, $h = 127$; and A. Ullrich, $h = 120$. It can be seen that, not surprisingly, all of these highly cited researchers also have high *h* indices and that high *h* indices in the life sciences are much higher than in physics. Among 36 new inductees in the National Academy of Sciences in biological and biomedical sciences in 2005, I find $\langle h \rangle = 57$, $\sigma_h = 22$, highest $h = 135$, lowest $h = 18$, and median $h_m = 57$. These latter results confirm that *h* indices in biological sciences tend to be higher than in physics; however, they also indicate that the difference appears to be much higher at the high end than on average. Clearly, more research in understanding similarities and differences of *h* index distributions in different fields of science would be of interest.

In summary, I have proposed an easily computable index, *h*, which gives an estimate of the importance, significance, and broad impact of a scientist's cumulative research contributions. I suggest that this index may provide a useful yardstick with which to compare, in an unbiased way, different individuals competing for the same resource when an important evaluation criterion is scientific achievement.

I am grateful to many colleagues in the University of California at San Diego Condensed Matter group and especially Ivan Schuller for stimulating discussions on these topics and encouragement to publish these ideas. I also thank the many readers who wrote with interesting comments since this paper was first posted at arXiv.org (6); the referees who made constructive suggestions, all of which led to improvements in the paper; and Travis Brooks and the SPIRES database administration for rapidly implementing the *h* index in their database.

[‡]This was first introduced in the SPIRES database.

1. Laherrere, J. & Sornette, D. (1998) *Eur. Phys. J. E Soft Matter* **B2,** 525–539.
2. Redner, S. (1998) *Eur. Phys. J. E Soft Matter* **B4,** 131–134.
3. Redner, S. (2005) *Phys. Today* **58,** 49–54.
4. van Raan, A. F. J. (2004) *Scientometrics* **59,** 467–472.
5. King, C. (2003) *Sci. Watch* **14,** no. 5, 1.
6. Hirsch, J. E. (2005) *arXiv.org E-Print Archive* (Aug. 3, 2005). Available at http://arxiv.org/abs/physics/0508025.

Downloaded at GRANADA/F CIENCIAS on August 25, 2020