Measurement, 10: 123–146, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366367.2012.711147

Psychology Press
Taylor & Francis Group

FOCUS ARTICLE

Measuring Scholarly Impact Using Modern Citation-Based Indices

John Ruscio
Psychology Department
The College of New Jersey

Florence Seaman
Department of Counseling and Clinical Psychology
Teacher's College, Columbia University

Carianne D'Oriano, Elena Stremlo, and Krista Mahalchik
Psychology Department
The College of New Jersey

Scholarly impact is studied frequently and used to make consequential decisions (e.g., hiring, tenure, promotion, research support, professional honors), and therefore it is important to measure it accurately. Developments in information technology and statistical methods provide promising new metrics to complement traditional information sources (e.g., peer reviews). The introduction of Hirsch's (2005) *h* index—the largest number *h* such that at least *h* articles are cited *h* times each, or the length of the largest square in a citations × articles array—sparked an explosion in research on the measurement of scholarly impact. We evaluate 22 metrics, including conventional measures, the *h* index, and many variations on the *h* theme. Our criteria encompass conceptual, empirical, and practical issues: ease of understanding, accuracy of calculation, effects on incentives, influence of extreme scores, and validity. Although the number of publications fares well on several criteria, the most attractive measures include *h*, several variations that credit citations outside the *h* square, and two variations that control for career stage. Additional data suggest that adjustments for self-citations or shared authorship probably would not improve these measures much, if at all. We close by considering which measures are most suitable for research and practical applications.

Keywords: citations, *h* index, scholarly impact, self-citation, shared authorship

Correspondence should be addressed to John Ruscio, Psychology Department, The College of New Jersey, Ewing, NJ 08628. E-mail: ruscio@tcnj.edu

How much impact has an individual's scholarship had on scientific theory, research, or practice? Scholarly impact is difficult to assess, yet it is the subject of empirical investigation as well as an important factor considered in many high-stakes decisions. For example, Duffy, Martin, Bryan, and Raque-Bogdan (2008) studied the impact of *Journal of Counseling Psychology* editorial board members; Smith (2010) studied the impact of scholars in the area of lesbian, gay, bisexual, and transgender scholarship; and Duffy, Jadidian, Webster, and Sandell (2011) studied the research productivity of 673 faculty affiliated with counseling and industrial-organizational psychology doctoral programs. Nosek et al. (2010) ranked not only the impact of social and personality psychologists, but also the programs to which they belong. How scholarly impact is assessed not only reflects but also has the potential to shape what scientists value and how they formulate their career goals. The balance between quantity and quality of work, or even the very kinds of research performed (e.g., theoretical vs. empirical projects, primary studies vs. meta-analyses, cross-sectional vs. longitudinal studies), can be influenced by the incentives created through a particular approach to assessing impact. Looking back rather than forward, historians of science might find metrics helpful tools to study the scholarly impact of influential figures. In addition to their increasing and potential uses in research, metrics are used for administrative purposes. A recent survey of 150 readers of *Nature* magazine asked scientists whether their institutions are using metrics to evaluate their scholarly impact (Abbott, Cyranoski, Jones, Maher, Schiermeier, & Van Noorden, 2010). Results suggest that metrics are commonly used to inform decisions regarding hiring, tenure, promotion, salary and bonuses, performance reviews and appraisals, and the allocation of research resources.

Whether for research or administrative purposes, it is essential that scholarly impact be measured as accurately and as fairly as possible. In recent years, developments in information technology and statistical methods have provided new tools to help assess scholarly impact. The emergence of user-friendly electronic databases cataloguing, not only publications, but also references and citations, has made it fairly easy to retrieve data on how frequently an individual's work has been cited. Using these data, Hirsch (2005) devised a simple, clever index of scholarly impact designed to reward both the quantity and quality of an individual's scholarship. Though citations have been counted for some time (e.g., Endler, Rushton, & Roediger, 1978), research on modern citation-based indices has blossomed in response to Hirsch's seminal work. Many investigators have proposed new measures—often variations on Hirsch's original theme—and discussed their conceptual merits (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009, and Panaretos & Malesios, 2009, review many facets of this literature). Commenting on this proliferation of metrics, Van Noorden (2010) suggests that the focus should shift from developing new measures to understanding and evaluating the merits of available measures. Our methodological investigation does precisely this.

There are significant challenges to evaluating scholarly impact accurately and fairly. Far too many scientists publish far too much research for anyone to be familiar with more than a small fraction. Moreover, research has become highly specialized and it is difficult to judge work outside one's areas of expertise. Due to the challenges posed by the sheer volume of individuals' research output and its specialized nature, it is often impractical for researchers or decision makers to carefully review this work. There are many qualitative indicators of scholarly impact, including external peer reviews, editing experience, membership on editorial boards, awards for scholarly achievement, fellow status or leadership in professional organizations, invited talks, and grant support. Abbott et al. (2010) spoke with provosts, department heads, and other administrators at research institutions around the world and report that these decision makers rely more heavily on these qualitative indicators than on metrics. External peer reviews are given considerable weight, but these are very demanding to produce and it may not be feasible to obtain them each time an important decision must be made. It might be even less feasible for researchers to process this wealth of information to compile a sufficiently large database of scholarly impact for study. In research and practice, quantitative measures are already being used to supplement qualitative indicators (Abbott et al., 2010; van Noorden, 2010).

Quantifying scholarly impact requires the use of heuristics, admittedly imperfect shortcuts that render the task manageable. Though there is an inevitable loss of information when it is summarized numerically, the use of transparent and objective measures can help to reduce the influence of biases that are less easily detected or eliminated when relying on qualitative indicators. For example, Abbott et al. (2010) note that the use of metrics can help to "break the old-boys' networks" that jeopardize the careers of women and minorities. Another potential advantage of using quantitative measures is that differences across disciplines and subdisciplines can be dealt with objectively. Publication and citation patterns differ widely across research areas (Hicks, 2006; Hirsch, 2005), and this poses an additional challenge for the assessment of scholarly impact. Unlike qualitative indicators, quantitative measures can be adjusted for discipline-specific publication and citation patterns (Petersen, Wang, & Stanley, 2010; Radicchi, Fortunato, & Castellano, 2008) or accompanied by empirically established norms to contextualize their meaning objectively. In the absence of a demonstrably appropriate adjustment, however, one should exercise considerable caution in comparing the scores of individuals working in disciplines with different publication norms (e.g., number of articles published, number of references cited per article).

Along similar lines, Hicks (2006) cautions that metrics can only be useful to the extent that what is considered to be impactful scholarly work in a discipline consists of publications catalogued in electronic databases. Hicks' review of pertinent evidence suggests that whereas psychology and economics may be sufficiently scientific in their publication patterns to support the use of quantitative measures, there is greater potential for incomplete and misleading results in sociology or other disciplines whose publication patterns more closely resemble the humanities. Our empirical data are drawn from the publications of psychologists, but we believe that our conclusions regarding the various strengths and weaknesses of the measures themselves will generalize to any sufficiently scientific discipline that supports their use.

We review and evaluate metrics that have been developed to meet the challenges of assessing scholarly impact in appropriate scientific disciplines. These measures draw upon other scholars' knowledge and areas of expertise. For example, when a manuscript is evaluated for publication, the action editor and reviewers are well equipped to judge the work carefully. Acceptance for publication therefore provides some degree of quality assurance, likewise, when another scholar cites this work in a peer-reviewed outlet that signifies some degree of impact. Though publication and citation are not perfect indicators of impact, they afford objective measures that enable users to cope with the high volume and specialization of scholarship. We focus our attention on peer-reviewed journal articles because they are so highly regarded in scientific disciplines. Many of the measures we describe could be adapted to incorporate other types of scholarly work (e.g., books, chapters, conference presentations) that may be more systematically catalogued in the future. Whether such adaptations would increase or decrease the utility of these measures remains an empirical question.

CRITERIA FOR EVALUATING MEASURES OF SCHOLARLY IMPACT

Before proceeding, we want to stress that we do not advocate reliance on metrics to the exclusion of qualitative indicators when making personnel decisions, awarding grants, or conferring professional honors. Not only do these metrics rely on the qualitative evaluations of peer reviewers and action editors, but also the goal of the present investigation is to identify measures that offer the most useful contributions to a multifaceted evaluation. In research contexts, it may be less feasible to obtain peer reviews or to collect data on other qualitative indicators and incorporate it into data analyses. Because the value of well-chosen metrics can be substantial in either context, we assess the merits of quantitative measures using five criteria that encompass conceptual, empirical, and practical issues:

1. **Ease of understanding**. Conceptually simpler measures should be preferred to those that are more complex. A well-chosen measure should be easy to explain, particularly if it will be used by or communicated to individuals with little or no training in measurement or other quantitative methods.
2. **Accuracy of calculation**. Measures that are easier to calculate accurately should be preferred to those that are more error-prone. Though the complexity of a calculation is of little concern in research settings, where even complex algorithms can be computerized and error checked, in practice it may be desirable to use metrics that can be calculated without special-purpose software.[1] Even more important than the complexity of a measure is the availability of the information that one must retrieve to calculate a score. Measures that require less information or are more robust to missing information (e.g., articles that are not retrieved due to data-entry errors in the database) should be preferred.
3. **Effects on incentives**. How one measures scholarly impact can create or reinforce the perception that a research strategy focused on the quantity, quality, or balance between quantity and quality of publications will be rewarded (Haslam & Laham, 2010). Users should carefully select metrics that are consistent with the desired incentives to avoid influencing researchers' strategies in unwanted ways.
4. **Influence of extreme scores**. Distributions of citations are heavy-tailed. Clauset, Shalizi, and Newman (2009) found that citation counts can be modeled well using a power law distribution, and the citation counts for the 48,692 articles of our Sample 1 (see below) ranged from 0 to 3,848, with an interquartile range (*IQR*) of 1 to 19, a skewness of 18.63, and an excess kurtosis of 853.34. Even if many or most extreme scores correspond to genuinely influential articles that merit a strong influence on measures of scholarly impact, others may stem from good fortune (e.g., citation cascades unrelated to research quality), unique insights leading to "one-hit wonder" careers, or unrepresentative collaborative efforts (e.g., when a graduate student becomes coauthor on a faculty mentor's research). It is unclear how many highly cited articles fall into each of these categories. Users must decide for themselves whether they prefer metrics that are robust or sensitive to the influence of extreme scores. For example, if one's goal is to measure sustained or programmatic impact, robustness to extreme scores should be valuable.

[1] We wrote a computer program that calculates the 22 metrics assessed in the present work; this is available on request.

5. **Validity**. Naturally, a more valid measure is preferable to one that is less valid, and this is arguably the most important of our five criteria. We will place special emphasis on validity when discussing the selection of metrics for applications.

With these criteria in mind, we now review 22 metrics that have been introduced.

QUANTITATIVE MEASURES OF SCHOLARLY IMPACT

Suppose that Professor X has published 18 peer-reviewed journal articles, the rank-ordered citation counts for these articles are {24, 18, 12, 8, 6, 5, 5, 4, 4, 3, 2, 2, 1, 1, 1, 0, 0, 0}, and the first of these articles is 10 years old (i.e., publishing age = 10). There are many metrics that summarize this information. We present 22 metrics below, with more formal definitions shown in the Appendix.

Conventional measures

Perhaps the oldest and most popular measure is the number of articles published (N_a). For Professor X, $N_a = 18$.

The total number of citations to an individual's articles (C) is another conventional metric.[2] For Professor X, $C = 96$. Recently, it has become easy to retrieve citation counts. For example, the PsycINFO database provides extensive coverage of scholarly journals in our discipline (psychology) and closely related fields, and when one performs a search the results include "times cited in database" for each article; summing these yields C. Whereas this measure can be conceived as the area spanned within an array of citations × articles, many other measures are analogous to distances rather than areas. Thus, we also calculated the square root

[2] In addition to counting articles or citations, journals' reputations for publishing influential work can be quantified using expert ratings or objective indices. For example, as part of their Excellence in Research for Australia project, the Australian Research Council rated more than 20,000 scholarly journals spanning the full range of academic disciplines (http://www.arc.gov.au/era/era_journal_list.htm). Another approach familiar to most scientists is the Journal Impact Factor (JIF). This is usually calculated as the mean citations that all articles in two years' volumes of a journal receive within a subsequent one-year period (e.g., the mean citations in 2012 for all articles the journal published in 2010 and 2011). Garfield (2006) provides historical context and reviews strengths and weaknesses of the JIF, which was originally created to help identify journals for inclusion in research databases. When used to evaluate the impact of specific individuals or articles, the JIF is a fairly crude measure. It is based on all articles published in a journal, not just those pertinent to a particular evaluation, and the mean is a poor measure of central tendency for heavy-tailed citation count distributions. Haslam and Koval (2010) recorded the citation counts accumulated over 10 years for each of the 1,580 articles published in 1998 in the Web of Science's social-personality journal group. Using linear regression, they examined many predictors of citation impact. The JIFs for the journals in which the articles appeared predicted only 30% of their 10-year citation impact. Clearly, there is more to scholarly impact than the JIF captures.

We chose not to study the JIF for a number of reasons. The JIF was designed to measure journal impact, not individuals' scholarly impact. It is by no means clear whether or how one could adapt it for the latter purpose. Whereas citation counts are available for all articles appearing in a database such as PsycINFO or Web of Science, the JIF is not: Many new and highly specialized journals do not have a JIF. Moreover, it is much more cumbersome to retrieve JIF values than article citation counts, particularly if the JIF for the year of each article's publication is used rather than simply the most recently released JIF for a given journal. Developing a JIF-based metric was not feasible.

of C (*sqrt-C*) to examine whether a distance-like variant of this measure might outperform the original (e.g., by exhibiting greater robustness to the influence of the extreme scores that occur so often in the heavy-tailed distributions of citation counts). For Professor X, $sqrt\text{-}C = 9.80$.

The total number of citations may not reflect sustained or programmatic impact. For example, a student might make minor contributions to a highly influential article coauthored with his or her mentor but never again publish an influential paper. For a number of years, including the critical early career period, this individual's total citation count could be higher than that of many others at the same stage who have actually done more original, high-quality work. Two related measures attempt to control for this type of anomaly by calculating either the mean or the median number of citations per article (M_c and Mdn_c, respectively). For Professor X, $M_c = 5.33$ and $Mdn_c = 3.50$.

Modern citation-based indices

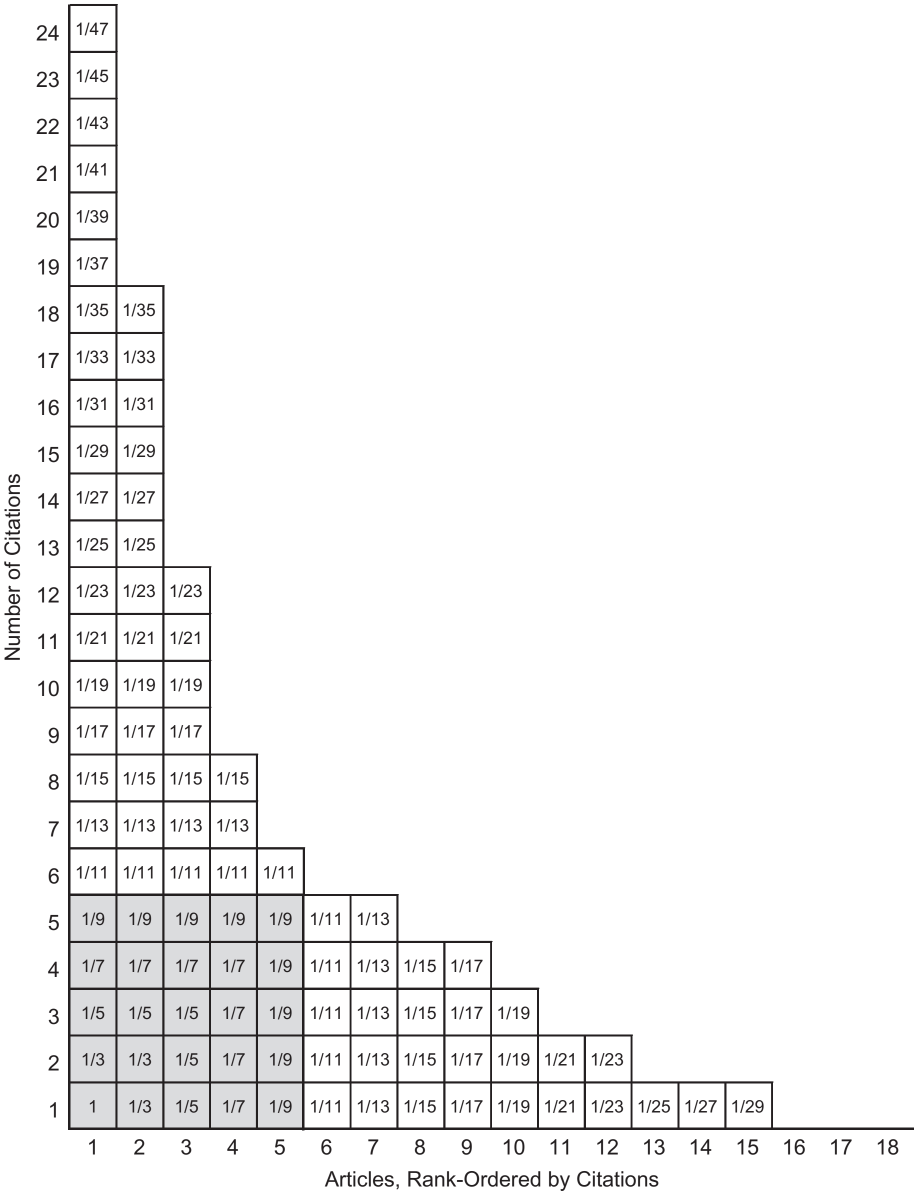
The h *index*

Hirsch (2005) introduced the h index, defined simply as the largest number h such that at least h articles are cited at least h times each. Professor X published 5 articles cited at least 5 times each, but not 6 articles cited at least 6 times each, so $h = 5$. Using the illustrative data for Professor X, Figure 1 plots the number of citations for each article to reveal a graphical representation of the h index: h equals the length of the largest square that fits within the resulting array of citations × articles.[3] Hirsch's h index sparked a great deal of research.

Crediting "excess" citations

Many proposed variations on the h theme award some credit for "excess" citations that fall outside the h square. For example, Professor X has a total of $C = 96$ citations, but because $h = 5$ this means that only 25 of the 96 citations "counted" in the calculation of the h index by falling inside the h square; 71 excess citations ($C - h^2 = 96 - 25 = 71$) received no credit. It is true that these excess citations do not yet count toward Professor X's score on h, but many of them will fall inside the h square if h increases in value. For example, if Professor X's sixth- or seventh-most-cited article were cited one more time, h would increase to 6, in which case 10 citations would no longer be counted as excess citations (with this one additional citation, $C - h^2$ would yield $97 - 36 = 61$, 10 fewer than the original 71 excess citations). Thus, among authors with equal scores on h, those with more excess citations are better positioned to increase their scores than those with fewer excess citations. Whereas a static view of h suggests that it ignores excess citations, a dynamic view reveals that excess citations are necessary for h to grow. Nonetheless, an index that awards credit for excess citations might prove more useful than h.

[3] More formally, h is the length of the Durfee square within a Ferrers diagram that partitions total citations by article (Andrews, 1984). We will refer to the h square rather than the Durfee square and a citation array rather than a Ferrers diagram.



24	1/47																	
23	1/45																	
22	1/43																	
21	1/41																	
20	1/39																	
19	1/37																	
18	1/35	1/35																
17	1/33	1/33																
16	1/31	1/31																
15	1/29	1/29																
14	1/27	1/27																
13	1/25	1/25																
12	1/23	1/23	1/23															
11	1/21	1/21	1/21															
10	1/19	1/19	1/19															
9	1/17	1/17	1/17															
8	1/15	1/15	1/15	1/15														
7	1/13	1/13	1/13	1/13														
6	1/11	1/11	1/11	1/11	1/11													
5	1/9	1/9	1/9	1/9	1/9	1/11	1/13											
4	1/7	1/7	1/7	1/7	1/9	1/11	1/13	1/15	1/17									
3	1/5	1/5	1/5	1/7	1/9	1/11	1/13	1/15	1/17	1/19								
2	1/3	1/3	1/5	1/7	1/9	1/11	1/13	1/15	1/17	1/19	1/21	1/23						
1	1	1/3	1/5	1/7	1/9	1/11	1/13	1/15	1/17	1/19	1/21	1/23	1/25	1/27	1/29			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

FIGURE 1 Illustrative citation data for Professor X. The largest square that fits within the array of citations (shaded in the graph) is 5 units in length, so $h = 5$ and the first 5 articles constitute the Hirsch core. Fractions represent credit assigned for each citation when calculating the h_t index; in this case, summing all fractional credit yields $h_t = 8.98$.

One such candidate is the tapered *h* index (h_t), which awards credit for every citation in an elegant manner that maintains a close connection to *h* (Anderson, Hankin, & Killworth, 2008). The fractions in the graph in Figure 1 reveal the algorithm by which credit decreases (tapers) for citations farther from the origin; h_t is the sum of the fractions for all citations. The first citation to an article is worth 1 point; with only this citation, both h_t and *h* would equal 1. Expanding this to a 2 × 2 square includes 3 additional citations worth 1/3 point each; summing points within this square yields $h_t = 1 + (3)(1/3) = 2$, again equal to *h*. Expanding this to a 3 × 3 square includes 5 additional citations worth 1/5 point each, summing to $h_t = 1 + (3)(1/3) + (5)(1/5) = 3 = h$. So far, it seems that *h* and h_t are identical. For Professor X, this is true up through a 5 × 5 square: The summed credit for these 25 citations yields $h_t = 1 + (3)(1/3) + (5)(1/5) + (7)(1/7) + (9)(1/9) = 5 = h$. The difference between *h* and h_t lies in the handling of excess citations, in this case the 71 citations falling outside the *h* square. These do not affect *h*, but h_t continues to award credit for them following the same tapered pattern. For Professor X, the credit for all 96 citations sums to $h_t = 8.98$. In all cases, $h_t \geq h$.

Another way to credit excess citations is to include them in a cumulative average across articles. One of the earliest alternatives to *h* takes the following form: The score on the *g* index (Egghe, 2006) is the largest value *g* such that the mean citations for the *g* most highly cited articles is at least *g*. As with *h*, there is a graphical representation of *g*. Instead of fitting the largest square completely inside the array of citations, excess citations above the square are allowed to offset missing citations within the square. For Professor X, a 9 × 9 square can be filled by moving 22 of the 26 excess citations above it to the inside. Expanding to a 10 × 10 square would not work because only 23 excess citations would be available to offset 35 missing spaces within the square. In this case, therefore, $g = 9$. Two other metrics, *f* and *t* (Tol, 2009), operate in the same way: Whereas *g* uses the arithmetic mean (the sum of *n* values divided by *n*) to cumulate citations, *f* uses the harmonic mean (the reciprocal of the arithmetic mean of the reciprocals of the *n* values) and *t* uses the geometric mean (the n^{th} root of the product of the *n* values.). For Professor X, $f = 7$ and $t = 8$. In all cases, $h \leq f \leq t \leq g$.

With some authors arguing that *h* may be too conservative and others arguing that *g* may be too liberal, Alonso, Cabrerizo, Herrera-Viedma, and Herrera (2010) proposed a hybrid. Their *hg* index is calculated as the geometric mean of *h* and *g*. For Professor X, $hg = 6.71$. In all cases, $h \leq hg \leq g$.

Another five variations on the *h* theme also allow some credit for excess citations above the *h* square. The first three are based on citations to all articles in the "Hirsch core," or the *h* most highly cited papers. The *a* index (Jin, 2006) is the mean of these articles' citation counts, the *m* index (m_i; Bornmann, Mutz, & Daniel, 2008) is the median of these articles' citation counts, and the *r* index (Jin, Liang, Rousseau, & Egghe, 2007) is the square root of these articles' total citation count. For Professor X, with 5 papers in the Hirsch core, $a = 13.60$, $m_i = 12.00$, and $r = 8.25$. Closely related to the *r* index is the h_w index (Egghe & Rousseau, 2008), which is the square root of the total number of citations for the r_0 most highly cited articles, with r_0 defined as the largest value such that the total number of citations for the r_0 most highly cited articles divided by *h* is no more than the number of citations for article r_0. For Professor X, $r_0 = 3$ and $h_w = 7.35$. Finally, Cabrerizo, Alonso, Herrera-Viedma, and Herrera (2010) defined the q^2 index as the geometric mean of *h* and m_i. For Professor X, $q^2 = 7.75$. In all cases, *a*, m_i, *r*, h_w, and q^2 must equal or exceed *h*.

A more direct approach to crediting excess citations was taken by Zhang (2009), who proposed the *e* index as the square root of all excess citations for articles in the Hirsch core. For Professor X, there were 68 citations to the 5 articles in the Hirsch core. Subtracting the 25 that contributed to *h* leaves 43 excess citations for these articles, the square root of which yields $e = 6.56$. Zhang introduced *e* as a complementary metric to use along with *h*, but we will see that data show them to be highly redundant.

Geometric variants

The graphical representation of the *h* index is the length of the largest square that fits within the array of citations. The maximum product index (*mp*; Kosmulski, 2007), a close cousin of *h*, is the area of the largest rectangle that fits inside the array. This is calculated as the maximum product of an article's rank (1 = most cited through N_a = least cited) and its citation count. For Professor X, $mp = 36$ because 3 articles were cited at least 12 times each (also, 9 articles were cited at least 4 times each) and no other product exceeds 36. Whereas *mp* corresponds to an area, *h* and many of its variants correspond to distances. We therefore introduce *sqrt-mp* as the square root of *mp*. As with the *sqrt-C* variant of the conventional measure *C*, we expected the square root transformation to increase robustness to extreme scores. For Professor X, $\textit{sqrt-mp} = 6.00$. In all cases, $h \leq \textit{sqrt-mp} \leq mp$. The *sqrt-C* measure also represents a geometric variant of *h* that credits excess citations. When all citations are arranged into a square, *sqrt-C* is its length; it can take fractional values when *C* itself is not a perfect square.

More stringent variants

The *h* index requires that each of *h* articles be cited at least *h* times. A more stringent variation is the $h^{(2)}$ index, which requires that each of $h^{(2)}$ articles be cited at least $[h^{(2)}]^2$ times (Kosmulski, 2006). For Professor X, there are 3 papers with at least 3^2 citations, but not 4 papers with at least 4^2 citations, hence $h^{(2)} = 3$. One could vary the exponent to obtain more or less stringent criteria (e.g., an $h^{(3)}$ index would require that at least $h^{(3)}$ articles be cited at least $[h^{(3)}]^3$ times each), and an infinite number of alternatives exist that includes fractional exponents. We evaluate only the $h^{(2)}$ index. In all cases, $h^{(2)} \leq h$.

Measures controlling for career stage

Because citations accumulate, it is impossible for *h* or any variation on the *h* theme to decrease over time. This gives an automatic advantage to senior investigators relative to their more junior colleagues that one may or may not want to measure. In addition to the *h* index, Hirsch (2005) introduced the *m* quotient (m_q), calculated as *h* divided by publishing age, which in turn is defined as the number of years since one's first article was published (see Jensen, Rouquier, & Croissant, 2009, for a critique and Nosek et al., 2010, for career stage metrics that use samples of researchers as a normative basis). To award some credit for excess citations, we introduce a tapered variation (m_{qt}) defined as h_t divided by publishing age. For Professor X, $m_q = 0.50$ and $m_{qt} = 0.90$. In all cases, $m_q \leq h$, $m_{qt} \leq h_t$, and $m_q \leq m_{qt}$.

ASSESSING MEASURES USING CONCEPTUAL AND PRACTICAL CRITERIA

In this section, we assess each measure using three conceptual and practical criteria: ease of understanding, accuracy of calculation, and effects on incentives.

Ease of understanding

Perhaps the simplest measure to understand is the total number of published articles (N_a). The total number of citations (C) is also easy to understand, followed closely by the mean number of citations per article (M_c), the median number of citations per article (Mdn_c), and the square root of total citations (*sqrt-C*). The immediate popularity of the h index attests to how effectively its simple definition lends itself to numerical and graphical illustrations. With the exception of the h_t index, the other variations on the h theme introduce elements that are somewhat to considerably more complex. We single out the h_t index as an exception because it elegantly extends the basic h concept in a way that we have found comparatively easy to explain to nonspecialists, especially with a diagram. Thus, we would rate the conventional measures plus h and h_t as the easiest metrics to understand.

Accuracy of calculation

Accuracy of calculation depends on both the complexity of a measure's algorithm and its robustness to missing information (e.g., articles that are not retrieved due to data entry errors in the database). Obtaining N_a is easy—citation database search results will indicate the number of articles identified—as is summing citation counts to yield C. The *sqrt-C* measure is simply the square root of C, M_c is just C divided by N_a, and finding Mdn_c requires only ranking the citation counts and locating the middle value (or, for an even N_a, calculating the mean of the two middle values). Though these conventional measures are all easy to calculate, they are at least somewhat sensitive to missing information. N_a is the most straightforwardly affected by missing articles. C and *sqrt-C* will be affected to varying degrees depending on the frequency with which the missing articles were cited, and the average-citation measures M_c and Mdn_c can exacerbate the influence of missing data, especially for individuals with few publications.

Among the modern citation-based indices, h is arguably the easiest to obtain because it requires little more than ranking the citation counts; among the ranked values, one locates the last one that equals or exceeds its position in the list. Is the highest citation count greater than or equal to 1? Is the second highest greater than or equal to 2? Is the third highest greater than or equal to 3? The final affirmative answer in this series yields the value of h. Anyone can follow these steps and no calculation is required. The Web of Science and Scopus databases even provide h on request. Because $h \leq N_a$, h is less affected by missing data. For example, Professor X published 18 articles, with $h = 5$. Whereas N_a would be reduced by 1 unit for each missing article, h would not be reduced at all unless there was so much missing data that there were no longer at least 5 articles cited at least five times apiece. This could not happen with only 1 or 2 missing articles, because there were originally 7 cited at least five times. There is only a 4% chance that Professor X's h index would be affected with 3 missing articles, and the chance of this happening surpasses 50% only when 7 (of the total of 18) articles are missing. This is an impressive robustness to randomly missing data. For similar reasons, all of the modern citation-based indices should be relatively robust to missing data.

The m_q index is simply h divided by publishing age, which means it's both easy to calculate and similarly robust to missing data. The $h^{(2)}$ index is calculated in much the same manner as h, but with a more stringent criterion at each serial position among the citation counts. This is a minor increase in complexity, and its stringency makes it even more robust to missing data.

Several other measures require an iterative series of computations. Calculating f, t, or g is similar to counting one's way to h, but rather than referring only to the citation count at serial position i, one calculates an average (harmonic, geometric, or arithmetic mean, respectively) of the i most highly cited articles and asks whether this average value equals or exceeds i. This is tedious to do by hand. The hg index is the geometric mean of h and g, so it requires calculating both. Four measures follow a different kind of two-step procedure. First h is calculated to determine how many articles are in the "Hirsch core," and then the a, m_i, and r indices are calculated as the mean, median, and square root of the sum of these h citation counts, respectively; the e index is calculated as described earlier. The q^2 index is the geometric mean of h and m_i, so it requires calculating both. The h_w index follows a two-step procedure that is very similar to that used to calculate the r index, but more tedious at the first step.

The remaining metrics involve arithmetic that can be cumbersome with even a modest number of citations. The mp index requires that one rank-order citation counts and then multiply these counts by their serial positions; mp equals the largest of these products, and $sqrt\text{-}mp$ is simply the square root of mp. The h_t index requires summing fractions as shown earlier. Unless one is extremely careful and persistent, a computer is required to calculate h_t accurately.

In short, only h, $h^{(2)}$, and m_q approach the ease with which conventional measures—especially N_a—can be calculated and are, therefore, most likely to be calculated accurately even without the use of specialized software. All of the modern citation-based indices should be fairly robust to missing data, providing accurate results despite imperfections in data sources and retrieval mechanisms.

Effects on incentives

The use of N_a as a metric can exert a strong influence on the incentives that shape researchers' activities by rewarding only the quantity of research and not its quality. Because a highly influential article receives no more credit than any other, researchers might be tempted to publish as much, or as quickly, as possible, even if the resulting work is less important or done less carefully. The use of C or $sqrt\text{-}C$ as a metric should solve that problem, but it might lead to others. With no explicit reward for the quantity of work produced, completing work in a timely manner may seem less urgent. The effects could be beneficial (e.g., investigators may address more important problems even if they require greater time and effort) or detrimental (e.g., individuals with perfectionistic tendencies might spend more than the optimal share of their effort checking data entry, running and rerunning analyses, or revising manuscripts). Using M_c or Mdn_c should attain some degree of balance between the rewards for quantity and quality of research, though there remains the problem that this could reduce the incentive to produce new work among those who have established a good score on an average-based measure. The publication of anything with low citation potential (e.g., errata, commentaries, obituaries) could pull down the average. In fact, for individuals who have attained a high citation average, publishing anything new becomes a gamble: The higher one's current average, the less likely it is that a new article will be cited enough times to maintain—let alone increase—the average.

In contrast to conventional measures, the *h* index establishes and maintains incentives to strike a balance between the quantity and quality of publications. Striving for quantity alone is not likely to yield a large *h* index, as a high volume of infrequently cited work corresponds to a long, thin array of citations into which a large square will not fit. Likewise, striving for quality alone is not likely to yield a large *h* index, as a low volume of highly cited work corresponds to a tall, narrow array of citations into which a large square will not fit. Thus, achieving a high score on *h* requires a large quantity of high-quality publications. Bartolucci (in press) proposed a decomposition of *h* into two components labeled impact and concentration, which highlights the need to be highly cited and to have these citations spread across a large number of publications to attain a high *h* index. The $h^{(2)}$ index maintains an emphasis on balance similar to that of the *h* index, but it applies a more stringent criterion.

Most variations on the *h* theme tend to reward quality more than quantity. This includes all of the metrics that award credit for only those excess citations that fall above the *h* square, namely f, t, g, hg, a, m_i, r, q^2, and e. Because the h_t, *mp*, and *sqrt-mp* indices provide equal rewards for excess citations falling above or to the right of the *h* square, there is no built-in reward for quantity or quality of publications. These indices are flexible in that they provide rewards for either research strategy.

ASSESSING MEASURES USING EMPIRICAL CRITERIA

Despite its size, the literature on measures of scholarly impact contains few studies empirically assessing metrics' performance. Alonso et al. (2009) describe four studies that compare the *h* index to conventional measures (e.g., N_a, C) and five studies that examine the correlations among metrics. A briefer review in Panaretos and Malesios (2009) partially overlaps with that of Alonso et al. Findings support the potential utility and incremental validity of *h* and some of its variations relative to conventional measures and reveal strong correlations between *h* and many of the proposed alternatives (e.g., Bornmann, Mutz, & Daniel, 2008; Bornmann, Wallon, & Ledin, 2008; Franceschet, 2009; Hirsch, 2007; Jensen et al., 2009). In the concluding remarks of their paper, Alonso et al. call for "deeper analyses and comparisons among all *h*-related indices" (2009, p. 286). We compiled a large data set that afforded rigorous tests of how extreme scores influence each metric as well as the metrics' validities.

Sample 1

We obtained a large sample of data representative of professors working at U. S. universities with doctoral programs in psychology. The National Research Council (NRC) ranked 185 such programs (Goldberger, Maher, & Flattau, 1995), and we sampled 10 professors at random from each of the 175 programs with departmental websites (or university directories) that listed faculty members' names and ranks. We recorded the score assigned to each department by the NRC, which ranged from 29 to 72 in an approximately normal distribution.

For each faculty member, we recorded his or her rank ($n = 450$ assistant professors, $n = 471$ associate professors, and $n = 829$ full professors); adjunct, affiliated, and emeritus faculty were excluded, and distinguished professors were assumed to have achieved the rank of full professor. Names were entered into advanced searches on PsycINFO with results restricted to peer-reviewed journal articles.[4] Individuals were excluded if their names did not appear in PsycINFO or if it was impossible to differentiate publications for authors with identical or nearly identical names. Many professors published under names that differed from listings on departmental websites (e.g., "Bob Smith" published as "Robert Smith"). We dropped middle names or initials unless they were necessary to differentiate between multiple authors with otherwise identical names; this increased the yield for many authors whose names appeared inconsistently in the PsycINFO database. We examined search results to ensure that only articles by the target author had been identified, regardless of the presence/absence of a middle name or initial. We cannot verify that we attained perfect sensitivity or specificity, but we did our best to approach these ideals within the constraints of the research team's available search time. Citation counts and publishing age were recorded for 1,750 professors who published 48,692 peer-reviewed journal articles that were cited 919,883 times. Table 1 presents the correlations between all 22 measures. Because metrics' distributions often differed substantially from one another, we include Spearman rank-order correlations in addition to the usual Pearson product-moment correlations.

Influence of extreme scores

We evaluated the influence of extreme scores on each measure in three ways: examining stability under repeated sampling, the change in scores when extreme values were added or removed, and the split-half reliability of each measure.

Stability

A measure that is more robust to extreme scores should be more stable under repeated sampling from a population than a measure that is more sensitive to extreme scores. The stability of each measure was tested in two ways. First, 100 bootstrap samples were drawn from the data (Efron & Tibshirani, 1993). For each of the 1,750 professors, the observed distribution of citation counts was treated as a population and a random sample of equal size was drawn with replacement. For each bootstrap sample, all 22 measures were calculated for all 1,750 professors. Then, scores on each measure were correlated across all pairwise combinations of the 100 bootstrap samples. The mean of these 4,950 correlations for each measure is listed in Table 2 (first column). A cluster of 6 modern citation-based indices exhibited the largest correlations ($r \geq .966$): h, h_t,

[4] In preliminary searches of psychological scientists, we found that PsycINFO identified as many or more articles and citations than Web of Science or Scopus. More recently, Ruscio and Prajapati (2012) found that the h indices of a sample of 286 university-affiliated psychology professors were extremely similar when calculated using the results of searches performed in the PsycINFO and Web of Science databases. Google Scholar results included conference presentations, unpublished manuscripts, and other types of sources among the target and citing works; we preferred to rely on peer-reviewed journal articles as citation targets. Readers interested in tools to perform Google Scholar searches, filter the results, and calculate citation indices can access two free programs: Scholarometer (http://scholarometer.indiana.edu/) and Publish or Perish (http://www.harzing.com/pop.htm).

TABLE 1
Correlations Between Measures of Scholarly Impact

	N_a	C	*sqrt-C*	M_c	Mdn_c	h	h_t	f	t	g	hg	a	m_i	r	h_w	q^2	e	mp	*sqrt-mp*	$h^{(2)}$	m_q	m_{qt}
N_{a}		.87	.87	.44	.33	.92	.91	.91	.91	.90	.92	.64	.63	.82	.81	.83	.75	.82	.82	.83	.51	.50
C	.76		1.00	.80	.65	.96	.98	.96	.98	.97	.97	.91	.87	.99	.99	.98	.97	.99	.99	.96	.64	.67
sqrt-C	.82	.90		.80	.65	.96	.98	.96	.98	.97	.97	.91	.87	.99	.99	.98	.97	.99	.99	.96	.64	.67
M_{c}	.24	.52	.62		.85	.68	.73	.68	.71	.72	.70	.93	.89	.85	.85	.82	.89	.84	.84	.78	.58	.64
Mdn_{c}	.07	.23	.31	.83		.60	.62	.59	.60	.57	.59	.68	.69	.68	.68	.68	.68	.67	.67	.65	.51	.54
h	.87	.85	.96	.48	.24		.99	.99	.99	.98	.99	.78	.76	.94	.94	.95	.89	.93	.93	.95	.66	.64
h_{t}	.87	.87	.98	.52	.26	.99		.99	1.00	.99	.99	.84	.82	.97	.96	.97	.93	.96	.96	.97	.65	.65
f	.86	.86	.97	.48	.23	.99	.99		.99	.98	.99	.79	.78	.94	.94	.95	.89	.93	.93	.95	.65	.66
t	.85	.87	.98	.50	.24	.99	1.00	1.00		.99	1.00	.82	.80	.96	.96	.96	.92	.95	.95	.96	.65	.65
g	.82	.89	.98	.52	.22	.97	.98	.98	.99		.99	.84	.81	.97	.96	.96	.93	.95	.95	.96	.65	.66
hg	.85	.88	.98	.50	.23	.99	1.00	.99	1.00	.99		.82	.79	.96	.95	.96	.92	.95	.95	.96	.66	.65
a	.52	.78	.86	.88	.54	.71	.76	.72	.75	.80	.76		.94	.94	.94	.91	.97	.94	.94	.86	.58	.65
m_{i}	.51	.67	.78	.85	.65	.67	.72	.68	.70	.72	.70	.91		.89	.90	.92	.92	.90	.90	.86	.57	.63
r	.76	.89	.99	.67	.33	.94	.97	.95	.96	.98	.97	.89	.80		1.00	.98	.99	.99	.99	.96	.66	.68
h_{w}	.75	.89	.99	.68	.34	.94	.96	.94	.96	.97	.96	.89	.81	1.00		.98	.99	.99	.99	.96	.65	.68
q^2	.78	.86	.98	.63	.35	.96	.98	.96	.97	.97	.97	.84	.83	.98	.98		.96	.98	.98	.97	.65	.68
e	.69	.88	.97	.72	.35	.89	.92	.90	.92	.95	.93	.93	.82	.99	.99	.96		.98	.98	.94	.64	.68
mp	.76	.98	.93	.60	.29	.88	.90	.89	.90	.91	.90	.83	.74	.93	.93	.90	.92		1.00	.95	.64	.67
sqrt-mp	.78	.86	.99	.67	.36	.94	.97	.95	.96	.96	.96	.88	.81	.99	.99	.98	.97	.91		.95	.64	.67
$h^{(2)}$	.73	.77	.95	.59	.30	.93	.94	.93	.94	.94	.94	.79	.75	.95	.95	.96	.93	.82	.95		.67	.68
m_{q}	.44	.47	.58	.35	.18	.60	.59	.60	.60	.58	.60	.46	.44	.59	.59	.61	.57	.50	.60	.62		.98
m_{qt}	.43	.48	.60	.41	.21	.58	.59	.58	.59	.58	.59	.51	.50	.61	.60	.62	.60	.52	.61	.63	.98	

Notes. Values on the diagonal (1.00) are omitted, values below the diagonal are the usual Pearson product-moment correlations, and values above the diagonal are Spearman rank-order correlations. Because values are rounded to 2 decimal places, an entry of 1.00 does not necessarily indicate perfect correlation.

f, *t*, *g*, and *hg*. Most of the remaining measures were also highly stable, with only 3 modern citation-based indices (a, m_{i}, and *mp*), M_{c}, and Mdn_{c} exhibiting $r < .90$.

The second way that we tested stability was to randomly sample citation counts for 28 articles (the mean number of articles per professor) from the distribution of all 48,692 articles. This yielded a set of scores on all measures, and the process was repeated 10,000 times. To calculate m_{q} and m_{qt}, which control for career stage, publishing age was held constant at 21 years (the mean for these professors). To examine stability under repeated sampling, the coefficient of variation ($CV = SD/M$) was calculated for each measure; these results are shown in Table 2 (second column). The same cluster of modern citation-based indices, joined by q^2, $h^{(2)}$, m_{q}, and m_{qt}, exhibited the greatest stability. Conventional citation measures C, M_{c}, and Mdn_{c} were considerably less stable, as were a and *mp*.

Adding or Removing Extreme Scores

A total of four extreme score analyses were performed. A Type I extreme score was constructed by doubling the largest citation count (e.g., if largest count was 25, an extreme score of 50 was added), and a Type II extreme score was constructed by adding a preceding digit of 1 to the largest citation count (e.g., if largest count was 25, an extreme score of 125 was added). The

TABLE 2
Influence of Extreme Scores on Measures of Scholarly Impact

	Stability Analyses		*Extreme Score Analyses: Median % Change*				
Index	*Mean r for 100 Bootstrap Samples*	*CV for 10,000 Randomly Sampled Scores*	*Type I Extreme Score Added*	*Type II Extreme Score Added*	*Type I + II Extreme Scores Added*	*Largest Value Dropped*	*Split-Half Reliability*
N_a	N/A	N/A	5.3	5.3	10.5	−5.3	N/A
C	.950	.494	53.4	118.5	179.5	−27.4	.954
sqrt-C	.963	.222	23.9	47.8	67.2	−14.8	.962
M_c	.791	.494	44.8	102.7	140.6	−22.0	.708
Mdn_c	.646	.457	9.1	9.1	20.0	−8.3	.586
h	.966	.154	7.7	7.7	16.7	−8.3	.980
h_t	.976	.138	13.6	16.8	29.9	−11.0	.984
f	.967	.142	10.0	10.0	20.0	−9.1	.977
t	.976	.141	12.5	15.8	28.6	−11.1	.986
g	.971	.188	19.1	31.2	46.1	−14.3	.982
hg	.975	.150	13.4	20.3	30.9	−11.7	.986
a	.839	.500	46.0	109.2	142.3	−22.5	.775
m_i	.700	.340	13.6	13.6	35.7	−11.7	.648
r	.951	.245	26.6	54.0	74.7	−16.3	.954
h_w	.944	.256	26.5	51.9	77.5	−15.5	.949
q^2	.929	.196	11.8	11.8	27.6	−11.0	.950
e	.928	.303	36.1	73.2	97.4	−20.6	.933
mp	.885	.697	33.3	138.5	190.2	−20.0	.919
sqrt-mp	.925	.277	15.5	54.4	70.3	−10.6	.934
$h^{(2)}$	.920	.155	0.0	0.0	20.0	0.0	.945
m_q	.919	.154	7.7	7.7	16.7	−8.3	.941
m_{qt}	.940	.138	13.6	16.8	29.9	−11.0	.946

Notes. *CV* = coefficient of variation (*SD*/*M*); Type I extreme score = doubled the largest citation count for each professor (e.g., if largest count was 25, added an extreme score of 50); Type II extreme score = added a preceding digit of 1 to the largest citation count for each professor (e.g., if largest count was 25, added an extreme score of 125).

median percent increase in each metric was calculated when adding a Type I extreme score, a Type II extreme score, or one of each. Additionally, the largest citation count was dropped for each professor and the median percent decrease in each metric was calculated. Results for these extreme score analyses are shown in Table 2 (four labeled columns). Among the most robust measures were N_a, h, and m_q, which increase with the number of extreme scores but not with their magnitude, and $h^{(2)}$, which does not necessarily increase at all due to its stringent criterion. Tapered credit for outlying citations muted the influence of extreme scores on h_t and m_{qt}. Mdn_c was also highly robust, as were modern citation-based indices f, t, hg, m_i, and q^2. The measures most sensitive to the influence of extreme scores were C, M_c, a, and *mp*.

Reliability

As a final test of the influence of extreme scores, we estimated each measure's reliability. Just as one can estimate the reliability of scores on a test by using the split-half method, we estimated the reliability of each measure in this way. PsycINFO lists articles chronologically, and we calculated each measure separately using odd- and even-numbered articles (e.g., for an author who published 10 articles, we calculated *h* for the subset of articles appearing in positions 1, 3, 5, 7, and 9 on the PsycINFO list as well as for articles appearing in positions 2, 4, 6, 8, and 10); 90 individuals with only one article were dropped from these analyses. This is analogous to splitting a test into halves using odd- and even-numbered items. We applied the usual Spearman-Brown correction of $r_{xx} = (2 \times r_{12}) / (1 + r_{12})$, where r_{12} is the correlation between halves and r_{xx} is the corrected estimate of reliability. Because extreme scores among citation counts should contribute differentially to the scores for odd- and even-numbered articles, reliability estimates should be larger for measures that are more robust to the influence of extreme scores. Reliability estimates are shown in Table 2 (final column). Split-half reliabilities exceeded .90 for most measures, but reliability was especially high ($r_{xx} > .97$) for the six modern citation-based indices that clustered together in each analysis so far. Reliability was quite a bit lower ($r_{xx} < .80$) for four measures (M_c, Mdn_c, *a*, and m_i) and intermediate for the remaining measures.

Validity

We assessed each measure on the final and most important of our five criteria in two ways. First, we calculated a standardized measure of effect size (ω^2) for the mean difference across professor ranks. All else being equal, one would expect a valid metric to reflect the greater scholarly impact of professors at higher ranks. We did not expect especially large values of ω^2 because we recognize that there are many other factors involved in attaining promotion and achieving scholarly impact. Nonetheless, this comparison affords some insight into metrics' relative validities. For all but m_q and m_{qt}, which are designed to control for career stage, means followed the expected order of full professors > associate professors > assistant professors. The magnitude of ω^2 served as our first measure of validity (see Table 3, first column).

Along with the conventional measure N_a, the modern citation-based indices *h*, h_t, *f*, *t*, *g*, and *hg* differentiated most strongly between professor ranks ($.178 \leq \omega^2 \leq .202$ for these metrics). Most of the other conventional measures fared comparatively poorly ($\omega^2 < .100$ for *C*, M_c, and Mdn_c; but $\omega^2 = .171$ for *sqrt-C*), as did two of the citation-based indices ($\omega^2 = .064$ for *a* and .055 for m_i). The extent to which the two metrics designed to control for career stage attained this goal is reflected in their very small differences across professor ranks ($\omega^2 = .007$ and .008 for m_q and m_{qt}, respectively). The remaining seven citation-based indices attained modest validity ($.105 \leq \omega^2 \leq .159$) in differentiating professor ranks.

As a second test of validity, we calculated correlations between each metric and NRC school scores. Here, too, we recognize the imperfection of the criterion (e.g., scores are based on whole psychology departments, not specific areas or programs, let alone individual faculty) but maintain that the pattern of correlations provides useful information about the metrics' relative validities. All correlations were positive, and their magnitudes served as our second measure of validity (see Table 3, second column).

The most striking trend in these results is that with only one exception, all of the modern citation-based indices yielded correlations that were larger (each $r > .26$) than those for conventional measures (each $r < .26$). The exception was the relatively large correlation ($r = .346$) for *sqrt-C*. The same cluster of metrics that was especially reliable and that differentiated most

TABLE 3
Validity Coefficients for Measures of Scholarly Impact

Index	*Effect size* (ω^2) *across Professor Ranks*	*Correlation with NRC School Scores*
N_a	.200	.226
C	.097	.258
sqrt-C	.171	.346
M_c	.017	.249
Mdn_c	.004	.126
h	.202	.335
h_t	.195	.339
f	.197	.333
t	.194	.340
g	.178	.342
hg	.192	.341
a	.064	.306
m_i	.055	.291
r	.150	.354
h_w	.147	.354
q^2	.156	.355
e	.120	.352
mp	.105	.281
sqrt-mp	.156	.351
$h^{(2)}$	.159	.366
m_q	.007	.269
m_{qt}	.008	.278

Notes. The sample of 1,750 professors included 450 assistant professors, 471 associate professors, and 829 full professors. NRC = National Research Council.

strongly between professor ranks—h, h_t, f, t, g, and hg—achieved relatively large correlations with school scores ($.333 \leq r \leq .342$). This was surpassed, by a small margin, by six other modern indices—r, h_w, q^2, e, *sqrt-mp*, and $h^{(2)}$ ($.351 \leq r \leq .366$). The two metrics designed to control for career stage yielded smaller correlations ($r = .269$ for m_q and .278 for m_{qt}), perhaps in part because "older" departments are, on average, more highly ranked than "younger" departments. The remaining three indices attained more moderate correlations ($r = .281$, .291, and .306 for mp, m_i, and a, respectively).

ADJUSTING FOR SELF-CITATIONS AND SHARED AUTHORSHIP

In our main study, we assessed measures of scholarly impact in their original form, unadjusted for self-citations or shared authorship. Many investigators argue that self-citations should be identified and removed (e.g., Schreiber, 2007; Zhivotovsky & Krutovsky, 2008) or that corrections should be made for shared authorship (e.g., Egghe, 2008; Schreiber, 2008, 2010) and that failing to do so compromises the utility of these measures. These criticisms are not self-evidently true, and we address them empirically.

There are many reasons why the works cited in a target article might include one or more of its authors (Costas, van Leeuwen, & Bordons, 2010). For example, self-citations can be appropriate (e.g., citing earlier stages in a program of research), gratuitous (e.g., attempts to inflate one's apparent record of achievement), convenient (e.g., authors might be more familiar with their own work than that of others), or incidental (e.g., a citation was included in the draft of a manuscript before one of the cited authors was invited to join as a coauthor). Identifying self-citations requires considerable effort. One needs to define what qualifies as a self-citation and inspect every citation to every article published by the target author to determine whether it meets this definition. We operationalized self-citations in five ways: (1) the first citing author matched the first cited author, (2) the first citing author matched any cited author, (3) the target author matched any cited author, (4) any citing author matched the first cited author, and (5) any citing author matched any cited author. Five variants of each metric were created by removing self-citations according to each of these criteria.

For shared authorship, we coded two variables—the number of authors and the target author's position on the list of authors—for each article and adjusted citations counts in four ways. Each of these divides credit such that the sum remains 100% across all authors: (1) evenly divided credit, (2) double credit for first author, (3) linearly declining credit, and (4) harmonically declining credit (Hagen, 2008). Four variants of each metric were created by adjusting citation counts in each of these ways.

Sample 2

To ensure that these adjustments would influence measures' values, we set a minimum of 20 peer-reviewed articles for an author to qualify for inclusion in this study. Within the author index on PsycINFO, we identified the first author whose last name began with "A" and who had at least 20 peer-reviewed articles. To ensure that we were collecting data on contemporary researchers, the author also had to have a publication within the past two years. Once these requirements were met, the author was added to the study and the search for the next author began with the letter "B," and so forth. PsycINFO searches were performed and data recorded as for Sample 1. Sample 2 included data for 156 authors who published 7,101 articles that were cited 91,265 times.

Removing self-citations

Across all 156 authors, the percent of citations coded as self-citations ranged from $Mdn = 5.23\%$ ($IQR = 3.04\%$ to 9.01%) for the most stringent criterion to $Mdn = 15.79\%$ ($IQR = 11.20\%$ to 22.04%) for the most liberal criterion. All rank-order correlations between the original (unadjusted) version of each metric and its five self-citation variants exceeded .950, and most (89%) exceeded .980; $Mdn = .992$. Given these extremely high correlations, it should not be surprising that split-half reliabilities changed very little after adjustments were made. In most instances (78%), there was a decrease in reliability, though all changes were quite small: No reliability changed by more than ±.03, and most reliabilities (87%) did not change by more than ±.01. Among the 22% of instances in which reliability increased slightly, no criterion of self-citation was consistently superior. Especially if the identification and removal of self-citations is labor intensive, the present evidence supports the use of unadjusted measures.[5]

Adjusting for shared authorship

Across all 7,101 articles in this sample, the norm was to have four coauthors ($Mdn = 4$, $IQR =$ 2 to 6). Across the 156 authors in this sample, the average number of coauthors on their work ranged from $Mdn = 1$ to 11, though most authors (90%) averaged six or fewer coauthors and the middle 50% of all authors ranged from $Mdn = 3$ to 5 co-authors. All four credit-sharing algorithms yielded median author credit values of about 30% per article ($IQR \approx 21\%$ to 44%).

Rank-order correlations between the original version of each metric and its adjusted variants ranged from .790 to .912 ($Mdn = .854$). In nearly all comparisons (96%), the original version of each metric was more reliable than its adjusted variant. Adjustments decreased reliability by as much as .073 (*Mdn* decrease = .036). The few increases were of small magnitude ($\leq .021$) and only occurred for the Mdn_c and t indices. These results provide no compelling reason to adjust for shared authorship.

Though adjusting for shared authorship may be intuitively appealing and one might construct alternative weighting schemes to make adjustments, we are not optimistic that any algorithm will handle the complexities of shared authorship very effectively. For example, whereas in psychology authorship is supposed to indicate the "relative scientific or professional contributions of the individuals involved" (American Psychological Association, 2002, section 8.12[b]), these guidelines are not always followed. Authorship is often awarded for activities that can be essential to completing a research project but that might involve little or no scientific or professional contribution (e.g., funding or managing a laboratory, providing access to existing records, assisting with data collection), which the APA recommends acknowledging in a footnote or opening statement rather than with authorship. Moreover, conventions for order of authorship can vary by discipline, which poses problems in adjusting for shared authorship when scientists work in different disciplines or publish interdisciplinary work.

SELECTING MEASURES FOR APPLICATIONS IN RESEARCH OR PRACTICE

Modern citation-based indices of scholarly impact have much to offer as research tools and as supplements to qualitative information in many important decision-making contexts. To inform the choice of measures, we assessed their strengths and weaknesses with respect to conceptual, empirical, and practical issues. We consider validity to be the most important criterion. Among the conventional measures, only N_a achieved a respectable degree of validity. It showed a strong difference across professor ranks, though a comparatively small correlation with NRC school scores. The conventional measures C, M_c, and Mdn_c fared more poorly, especially across professor ranks. Though not strictly a conventional measure, *sqrt-C* performed well, presumably

[5] PsycINFO does not identify or remove self-citations, but the Web of Science and Scopus databases do. At a user's request, they remove self-citations by applying our most liberal operationalization of self-citation (any citing author matches any cited author).

because the square root transformation reduced the influence of extreme scores (the same was true for the performance of *sqrt-mp* relative to the original *mp* index). Many of the modern citation-based indices, in contrast, exhibited good validity in both tests. Among the most valid were the *h* index and several close variants, including h_t, *f*, *t*, *g*, and *hg*. These indices also were remarkably highly correlated; no Pearson or Spearman correlation among these measures fell below .97. This means that they would yield very similar results in research (e.g., comparable correlations with other variables) or practice (e.g., comparable rank-ordering of individuals). To help users choose among these indices, we offer recommendations based on the context of application.

The easiest of the valid measures to understand and calculate is the *h* index. This might be especially desirable when the audience includes nonspecialists, such as academic administrators or substantive researchers more interested in findings than methods. If the frequency of tied scores is a concern, the h_t index avoids this problem. For the 1,750 professors in Sample 1, there were only 49 unique scores on *h* but 1,568 unique scores on h_t. If fine distinctions are important, one might use h_t in place of *h* or use scores on h_t to break ties between individuals with equal scores on *h*. These indices' validities were similar, but h_t is less easy to understand or calculate accurately than *h*.

Another potentially important factor to consider is that some indices differentially reward the quantity or the quality of publications. The *h* index rewards a balance between quantity and quality because it corresponds to the length of the largest square that fits within a citation plot. Neither a few highly influential papers nor a large number of poorly-cited papers will produce a large *h*. If one prefers to reward quality more than quantity, one could use the *g*, h_t, or *sqrt-mp* indices. The *g* index awards credit for excess citations above the *h* square and is more sensitive to especially highly-cited papers than *h* or close cousins of *g*, namely *f* and *t*. The h_t index also awards credit for excess citations above the *h* square, and it was of comparable validity to *h* and *g*. The *sqrt-mp* index corresponds to the square root of the largest rectangle that fits within the array of citations × articles. This rectangle can be large if there are a handful of articles with very large citation counts. If one prefers to reward quantity more quality, one could use the h_t or *sqrt-mp* indices, which award credit for excess citations to the right of the *h* square in the same way as for excess citations above the square, or the conventional measure N_a, which is extremely easy to understand and calculate and was reasonably valid. Finally, one might prefer not to favor any particular research strategy. The h_t and *sqrt-mp* indices can be conceived as flexible variations on *h* in that they neither rigidly reward a balance between quantity and quality, as does *h*, nor favor quantity over quality (or vice versa). For example, if one investigator publishes 3 articles cited 50 times each and another publishes 50 articles cited 3 times each, neither would outscore the other on the h_t or *sqrt-mp* indices.

Though they did not appear among the most-valid indices, the two metrics designed to control for career stage, m_q and m_{qt}, performed fairly well. Their correlations with NRC school scores were reasonably strong. At least as important is that neither varied much across professor ranks, which suggests that some measure of control for career stage was achieved. Because they exhibited similar validities, the choice between them could be made in much the same way as the choice between *h* and h_t. Recall that m_q and m_{qt} are simply *h* and h_t divided by publishing age and the parallels become clear. Whereas m_q and *h* reward a balance between quantity and quality of publications, m_{qt} and h_t can flexibly reward either a greater quantity or a higher quality of publications.

The use of one or more well-chosen metrics based on readily available citation counts, as opposed to more laborious coding of inconsistently available qualitative data, should greatly facilitate research on scholarly impact. In practice and depending on the decision at hand, one might consider qualitative indicators such as candidates' editing experience, membership on editorial boards, awards for scholarly achievement, fellow status or leadership in professional organizations, invited talks, grant support, peer reviews, or other pertinent information sources. There is no flawless measure of scholarly impact, and impact itself is not always indicative of the most meritorious work (Sternberg, 2003; see especially the chapter by D. N. Robinson and the afterward by R. J. Sternberg). Nonetheless, consequential decisions must be made and evaluating an individual's achievements in multiple ways—including one or more metrics—should avoid the pitfalls of relying too heavily on any single mode of assessment.

For those intrigued by the possibilities afforded by using modern citation-based indices of scholarly impact in research or practice, note that the individual researcher need not be the unit of analysis. The *h* index and related measures have been used to rank the scholarly impact of groups of scientists (e.g., research groups, departments, universities, or countries), journals, and research topics (Alonso et al., 2009). For example, Nosek et al. (2010) used this methodology to rank the impact of programs in social and personality psychology (and their members), and Guerin (2010) used it to examine the impact of New Jersey colleges and universities on psychological science. When the individual is not the unit of analysis, one can either use the *h* index or a variation that adjusts for the number of articles published by each observational unit (see Molinari & Molinari, 2008). Developing and evaluating metrics affords many opportunities for individuals with expertise in measurement to contribute to the study of scholarly impact as well as applications in important decision-making contexts. In conclusion, to evaluate individuals' scholarly impact we believe that in most cases Hirsch's (2005) *h* index should be seriously considered.

REFERENCES

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & Van Noorden, R. (2010, June 17). Do metrics matter? *Nature*, *465*, 860–862. doi:10.1038/465860a

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). *h*-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, *3*, 273–289. doi:10.1016/j.joi.2009.04.001

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2010). *hg*-index: A new index to characterize the scientific output of researchers based on the *h*- and *g*- indices. *Scientometrics*, *82*, 391–400. doi:10.1007/s11192-009-0047-5

American Psychological Association. (2002). American Psychological Association ethical principles of psychologists and code of conduct. Retrieved from http://www.apa.org/ethics/code2002.html

Anderson, T., Hankin, K., & Killworth, P. (2008). Beyond the Durfee square: Enhancing the *h*-index to score total publication output. *Scientometrics*,*76*, 577–588. doi:10.1007/s11192-007-2071-2

Andrews, G. E. (1984). *The theory of partitions*. New York, NY: Cambridge University Press.

Bartolucci, F. (in press). On a possible decomposition of the h-index. *Journal of the American Society for Information Science and Technology*.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2008). Are there better indices for evaluation purposes than the *h* index? A comparison of nine different variants on the *h* index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, *59*, 830–837. doi:10.1002/asi.20806

Bornmann, L., Wallon, G., & Ledin, A. (2008). Is the *h*-index related to (standard) measures and to the assessments by peers? An investigation of the *h*-index by using molecular life sciences data. *Research Evaluation*, *17*, 149–156. doi:10.3152/095820208X319166

Cabrerizo, F. J., Alonso, S., Herrera-Viedma, E., & Herrera, F. (2010). q^2-index: Quantitative and qualitative evaluation based on the number and impact of authors in the Hirsch core. *Journal of Informetrics*, *4*, 23–28. doi:10.1016/j.joi.2009.06.005

Clauset, A., Shalizi, C. S., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*, 661–703. doi:10.1137/070710111

Costas, R., van Leeuwen, T. N., & Bordons, M. (2010). Self-citations at the meso and individual levels: Effects of different calculation methods. *Scientometrics*, *82*, 517–537. doi:10.1007/s11192-010-0187-7

Duffy, R. D., Jadidian, A., Webster, G. D., & Sandell, K. J. (2011). The research productivity of academic psychologists: Assessment, trends, and best practice recommendations. *Scientometrics*, *89*, 207–227. doi:10.1007/s11192-011-0452-4

Duffy, R. D., Martin, H. M., Bryan, N. A., & Raque-Bogdan, T. L. (2008). Measuring individual research productivity: A review and development of the Integrated Research Productivity Index. *Journal of Counseling Psychology*, *55*, 518–527. doi:10.1037/a0013618

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. San Francisco, CA: Chapman & Hall.

Egghe, L. (2006). Theory and practise of the *g*-index. *Scientometrics*, *69*, 131–152. doi:10.1007/s11192-006-0144-7

Egghe, L. (2008). Mathematical theory of the *h*- and *g*-index in case of fractional counting of authorship. *Journal of the American Society for Information Science and Technology*, *59*, 1608–1616. doi:10.1007/s11192-006-0144-7

Egghe, L., & Rousseau, R. (2008). An *h*-index weighted by citation impact. *Information Processing and Management*, *44*, 770–780. doi:10.1016/j.ipm.2007.05.003

Endler, N. S., Rushton, J. P., & Roediger, H. L., III (1978). Productivity and scholarly impact (citations) of British, Canadian, and U. S. departments of psychology (1975). *American Psychologist*, *33*, 1064–1082. doi:10.1037/0003-066X.33.12.1064

Franceschet, M. (2009). A cluster analysis of scholar and journal bibliometric indicators. *Journal of the American Society for Information Science and Technology*, *60*, 1950–1964. doi:10.1002/asi.21152

Garfield, E. (2006). The history and meaning of the Journal Impact Factor. *Journal of the American Medical Association*, *295*, 90–93. doi:10.1001/jama.295.1.90

Goldberger, M. L., Maher, B. A., & Flattau, P. E. (Eds.) (1995). *Research doctorate programs in the United States: Continuity and change*. Washington, DC: National Academies Press.

Guerin, M. T. (2010). Ranking the scholarly impact of New Jersey colleges and universities in the field of psychology. *TCNJ Journal of Student Scholarship*, *12*, 1–11.

Hagen, N. T. (2008). Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis. *PLoS One*, *3*(12), e4021. doi:10.1371/journal.pone.0004021

Haslam, N., & Koval, P. (2010). Predicting long-term citation impact of articles in personality and social psychology. *Psychological Reports*, *106*, 891–900. doi:10.2466/pr0.106.3.891.900

Haslam, N., & Laham, S. M. (2010). Quality, quantity, and impact in academic publication. *European Journal of Social Psychology*, *40*, 216–220. doi:10.1002/ejsp.727

Hicks, D. (2006). The dangers of partial bibliometric evaluation in the Social Sciences. *Economia Politica*, *23*, 145–162. doi:10.1428/22461

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*, 16569–16572. doi:10.1073/pnas.0507655102

Hirsch, J. E. (2007). Does the *h*-index have predictive power? *Proceedings of the National Academy of Sciences*, *104*, 19193–19198. doi:10.1073/pnas.0707962104

Jensen, P., Rouquier, J.-B., & Croissant, Y. (2009). Testing bibliometric indicators by their predictions of scientists' promotions. *Scientometrics*, *78*, 467–479. doi:10.1007/s11192-007-2014-3

Jin, B. (2006). *H* index: An evaluation indicator proposed by Scientist. *Science Focus*, *1*, 8–9.

Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The *R*- and *AR*- indices: Complementing the *h*-index. *Chinese Science Bulletin*, *52*, 855–863.

Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original *h*-index. *ISSI Newsletter*, *2*, 4–6.

Kosmulski, M. (2007). MAXPROD: A new index for assessment of the scientific output of an individual, and a comparison with the *h*-index. *International Journal of Scientometrics, Informetrics, and Bibliometrics, 11*. Retrieved from http://cybermetrics.cindoc.csic.es/articles/v11i1p5.pdf

Molinari, J. F., & Molinari, A. (2008). A new methodology for ranking scientific institutions. *Scientometrics*, *75*, 163–174. doi:10.1007/s11192-007-1853-2

Nosek, B. A., Graham, J., Lindner, N. M., Kesebir, S., Hawkins, C. B., Hahn, C., . . . Tenney, E. R. (2010). Cumulative and career-stage citation impact of social-personality psychology programs and their members. *Personality and Social Psychology Bulletin*, *36*, 1283–1300. doi:10.1177/0146167210378111

Panaretos, J., & Malesios, C. (2009). Assessing scientific research performance and impact with single indices. *Scientometrics*, *81*, 635–670. doi:10.1007/s11192-008-2174-9

Petersen, A. M., Wang, F., & Stanley, H. E. (2010). Methods for measuring the citations and productivity of scientists over time and discipline. *Physical Review E*, *81*, 036114. doi:10.1103/PhysRevE.81.036114

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*, 17268–17272. doi:10.1073/pnas.0806977105

Ruscio, J., & Prajapati, B. (2012). *Comparing scholarly impact between the PsycINFO and Web of Science databases using the h and m_q indices*. Manuscript in preparation.

Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *Euro-Physics Letters*, *78*, 30002. doi:10.1209/0295-5075/78/30002

Schreiber, M. (2008). A modification of the *h*-index: The h_{m}-index accounts for multi-authored manuscripts. *Journal of Informetrics*, *2*, 211–216.

Schreiber, M. (2010). A case study of the modified *g* index: Counting multi-author publications fractionally. *Journal of Informetrics*, *4*, 636–643. doi:10.1016/j.joi.2009.06.003

Smith, N. G. (2010). Productivity in lesbian, gay, bisexual, and transgender scholarship in counseling psychology: Institutional and individual ratings for 1990 through 2008. *Counseling Psychologist*, *38*, 50–68. doi:10.1177/0011000009345533

Sternberg, R. J. (Ed.). (2003). *The anatomy of impact: What makes the great works of psychology great*. Washington, DC: American Psychological Association.

Tol, R. S. J. (2009). The *h*-index and its alternatives: An application to the 100 most prolific economists. *Scientometrics*, *80*, 317–324. doi:10.1007/s11192-008-2079-7

Van Noorden, R. (2010, June 17). A profusion of measures. *Nature*, *465*, 864–866. doi:10.1038/465864a

Zhang, C.-T. (2009). The *e*-index, complementing the *h*-index for excess citations. *PLoS One*, *4*(5), e5429. doi:10.1371/journal.pone.0005429

Zhivotovsky, L. A., & Krutovsky, K. V. (2008). Self-citation can inflate *h*-index. *Scientometrics*, *77*, 373–375. doi:10.1007/s11192-006-1716-2

APPENDIX

Calculating Measures of Scholarly Impact

In the definitions presented below, we assume that the *i* articles are rank-ordered by citation count from highest to lowest.

Measure	*Definition*
N_{a}	number of articles
C	$\sum_{i=1}^{N_a} c_i$, where c_{i} is the citation count for article i
sqrt-C	$\sqrt{C}$
M_{c}	C/N_{a}

Mdn_{c}	if N_{a} is odd: c_{j}, where $j = (1 + N_{\mathrm{a}})/2$;
	if N_{a} is even: $(c_{\mathrm{j}} + c_{\mathrm{j+1}})/2$, where $j = N_{\mathrm{a}}/2$

h	largest value of i such that $c_{\mathrm{i}} \geq i$
h_{t}	$\sum_{i=1}^{N_1} \sum_{j=1}^{c_i} f_{ij}$, where N_1 is the number of articles cited at least once and $f_{\mathrm{ij}} = 1/\ (2 \times \max(i, j) - 1)$
f	largest value of j such that $\sqrt[j]{\prod_{i=1}^{j} c_i} \geq j$
t	largest value of j such that $\left(1 \Big/ \sum_{i=1}^{j} \frac{1}{c_i}\right) \geq j$
g	largest value of j such that $\left(\sum_{i=1}^{j} c_i \Big/ j\right) \geq j$
hg	$\sqrt{h \times g}$
a	$\sum_{i=1}^{h} c_i \Big/ h$

m_{i}	if h is odd: c_{j}, where $j = (1 + h)/2$
	if h is even: $(c_{\mathrm{j}} + c_{\mathrm{j+1}})\ /\ 2$, where $j = h/2$

r	$\sqrt{\sum_{i=1}^{h} c_i}$
h_{w}	$\sqrt{\sum_{i=1}^{r_0} c_i}$, where $r_0 =$ largest value of j such that $\left(\sum_{i=1}^{j} c_i \Big/ h\right) \leq c_j$
q^2	$\sqrt{h \times m_{\mathrm{i}}}$
e	$\sqrt{\left(\sum_{i=1}^{h} c_i\right) - h^2}$
mp	$\max(i \times c_{\mathrm{i}})$ for i from 1 to N_{a}
sqrt-mp	$\sqrt{mp}$
$h^{(2)}$	largest value of i such that $c_{\mathrm{i}} \geq i^2$
m_{q}	h/A, where A is publishing age (current year − year first article was published)
m_{qt}	h_{t}/A