Relative Citation Ratio (RCR): A new metric that uses citation rates to measure influence at the article level

B. Ian Hutchins [1], Xin Yuan [1], James M. Anderson [2] & George M. Santangelo [1]
[1] Office of Portfolio Analysis, [2] Division of Program Coordination, Planning, and Strategic Initiatives, National Institutes of Health, Bethesda, MD

Abstract

Despite their recognized limitations, bibliometric assessments of scientific productivity have been widely adopted. We describe here an improved method that makes novel use of the co-citation network of each article to field-normalize the number of citations it has received. The resulting Relative Citation Ratio is article-level and field-independent, and provides an alternative to the invalid practice of using Journal Impact Factors to identify influential papers. To illustrate one application of our method, we analyzed 88,835 articles published between 2003 and 2010, and found that the National Institutes of Health awardees who authored those papers occupy relatively stable positions of influence across all disciplines. We demonstrate that the values generated by this method strongly correlate with the opinions of subject matter experts in biomedical research, and suggest that the same approach should be generally applicable to articles published in all areas of science. A beta version of *iCite*, our web tool for calculating Relative Citation Ratios of articles listed in PubMed, is available at https://icite.od.nih.gov .

Introduction

In the current highly competitive pursuit of research positions and funding support (Couzin-Frankel, 2013), faculty hiring committees and grant review panels must make difficult predictions about the likelihood of future scientific success. Traditionally, these judgments have largely depended on recommendations by peers, informal interactions, and other subjective criteria. In recent years, decision-makers have increasingly turned to numerical approaches such as counting first or corresponding author publications, using the impact factor of the journals in which those publications appear, and computing Hirsch or H-index (Hirsch, 2005). The widespread adoption of these metrics, and the recognition that they are inadequate (Seglen, 1997; Anon, 2005, Anon, 2013), highlight the ongoing need for alternative methods that can provide effectively normalized and reliable data-driven input to administrative decision-making, both as a means of sorting through large pools of qualified candidates, and as a way to help combat implicit bias. A return to purely subjective evaluation with its attendant risk of partiality is neither desirable nor practical, and the use of metrics that are of limited value in decision-making is widespread and growing (Pulverer, 2013). The need for useful metrics is particularly pressing for funding agencies making policy decisions based upon the evaluation of large portfolios that often encompass diverse areas of science.

Though each of the above mentioned methods of quantitation has strengths, accompanying weaknesses limit their utility. Counting first or corresponding author publications does on some level reflect the extent of a scientist's contribution to their field, but it has the unavoidable effect of privileging quantity over quality, and may undervalue collaborative science (Stallings et al., 2013). Journal impact factor (JIF) was for a time seen as a valuable indicator of scientific quality because it serves as a convenient, and not wholly inaccurate, proxy for expert opinion (Garfield, 2006). However, its blanket use also camouflages large differences in the influence of individual papers. This is because impact factor is calculated as the average number of times articles published over a two-year period in a given journal are cited; in reality, citations follow a log-normal rather than a Gaussian distribution (Price, 1976; Wang et al., 2013). Moreover, since practitioners in disparate fields have differential access to high-profile publication venues, impact factor is of

limited use in multidisciplinary science-of-science analyses. Despite these serious flaws, JIF continues to have a large effect on funding and hiring decisions (Anon, 2005; Johnston, 2013; Misteli, 2013). H-index, which attempts to assess the cumulative impact of the work done by an individual scientist, disadvantages early career stage investigators; it also undervalues some fields of research by failing to normalize raw citation counts (Pulverer, 2013).

Alternative models for quantifying scientific accomplishment have been proposed but have not been widely adopted, perhaps because they are overly complicated to calculate and/or are difficult to interpret (Bollen et al., 2009; Waltman et al., 2011a). Some have dramatically improved our theoretical understanding of citation dynamics (Walker et al., 2007; Radicchi et al., 2008; Stringer et al., 2010; Wang et al., 2013). However, to combine a further technical advance with a high likelihood of widespread adoption by varied stakeholders, including scientists, administrators and funding agencies, several practical challenges must be overcome. Citation metrics must be article-level, field-normalized in a way that is scalable from small to large portfolios without introducing significant bias at any level, benchmarked to peer performance in order to be interpretable, and correlated with expert opinion. In addition, metrics should be freely accessible and calculated in a transparent way. Many efforts have been made to fulfill one or more of these requirements, including citation normalization to journals or journal categories (Moed et al., 1985; Zitt and Small, 2008; Opthof and Leydesdorff, 2010; van Raan et al., 2010; Waltman et al., 2011a, 2011b; Bornmann and Leydesdorff, 2013), citation percentiles (Bornmann and Leydesdorff, 2013; Bornmann and Marx, 2013), eigenvector normalization (Bergstrom and West, 2008; Bergstrom et al., 2008) or source-normalization (Zitt and Small, 2008; Moed, 2010) including the Mean Normalized Citation Score (Waltman et al., 2011a) and Source-Normalized Impact per Paper metrics (Moed, 2010). While all are improvements on Impact Factor, none meet all of the criteria listed above. Furthermore, these existing approaches are often unhelpful to decision-makers because they aggregate works from researchers across disparate geographical regions and institutional types. For example, current methods do not provide a way for a primarily undergraduate institutions to compare their portfolios against other teaching-focused institutions, nor do they allow developing nations to compare their research to that done in other developing nations (Crous, 2014). Incorporating a customizable benchmark as an integral part of an ideal citation metric would enable such an apples to apples comparison and facilitate downstream decision making activity.

We report here the development and validation of the Relative Citation Ratio (RCR) metric, which meets all of the above criteria and is based upon the novel idea of using the co-citation network of each article to field- and time-normalize by calculating the expected citation rate from the aggregate citation behavior of a topically linked cohort. An average citation rate is computed for the network, benchmarked to peer performance, and used as the RCR denominator; as is true of other bibliometrics, article citation rate (ACR) is used as the numerator. We use the RCR metric here to determine the extent to which National Institutes of Health (NIH) awardees maintain high or low levels of influence on their respective fields of research.

Results

Co-citation networks represent an article's area of influence

Choosing to cite is the long-standing way in which one scholar acknowledges the relevance of another's work. Before now, however, the utility of citations as a metric for quantifying influence has been limited, primarily because it is difficult to compare the value of one citation to another; different fields have different citation behaviors and are composed of widely varying numbers of potential citers (Jeong et al., 2003; Radicchi and Castellano, 2012). An effective citation-based evaluative tool must also take into account the length of time a paper has been available to potential citers, since a recently published article has had less time to accumulate citations than an older one. Finally, fair comparison is complicated by the fact that an author's choice of which work to cite is not random; a widely known paper is more likely to be referenced than an obscure one of equal relevance. This is because the accrual of citations follows a power law or log-normal pattern, in accordance with a process called preferential attachment (Jeong et al., 2003; Eom and

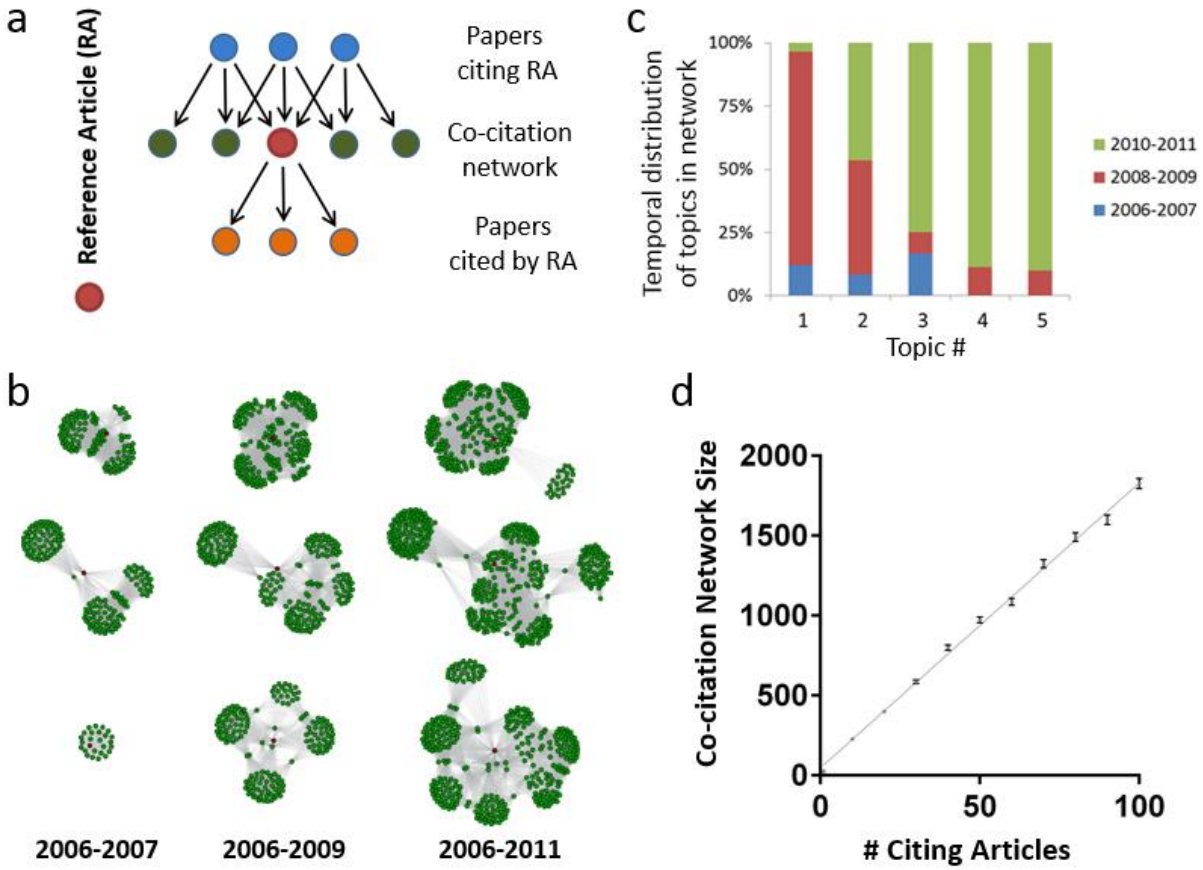


Figure 1. Properties of co-citation networks. (***a***) Schematic of a co-citation network. The Reference Article (RA) (red, middle row) cites previous papers from the literature (orange, bottom row); subsequent papers cite the RA (blue, top row). The co-citation network is the set of papers that appear alongside the article in the subsequent citing papers (green, middle row). The Field Citation Rate is calculated as the mean of the latter articles' journal citation rates. (***b***) Growth of co-citation networks over time. Three RAs published in 2006 (red dots) were cited 5 (top row), 9 (middle row), or 31 times (bottom row) by 2011. Three intervals were chosen to illustrate the growth of the corresponding co-citation networks: 2006-2007, 2006-2009, and 2006-2011 (the first, second, and third columns, respectively). Each article in one of the three co-citation networks is shown as a separate green dot; the edges (connections between dots) indicates their presence together in the same reference list. (***c***) Cluster algorithm-based content analysis of the 215 papers in the co-citation network of a sample reference article (RA; panel *b*, bottom network series) identified a changing pattern of relevance to different sub-disciplines over time. This RA described the identification of new peptides of possible clinical utility due to their similarity to known conotoxins. Papers in the co-citation network of this RA focused on: (1) α-conotoxin mechanisms of action; (2) structure and evolution of conotoxins; (3) cyclotide biochemistry; (4) conotoxin phylogenetics; and (5) identification and synthesis of lantibiotics. (***d***) Growth of an article's co-citation network is proportional to the number of times it has been cited. Each point is the average network size of 1000 randomly chosen papers with between 1 and 100 citations (error bars represent the standard error of the mean). Each paper is only counted once, even if it is co-cited with the article of interest multiple times. An average of 17.8 new papers is added to the co-citation network for each additional citation. This suggests substantial duplication of articles within a co-citation network, since on average 32.4 papers (median of 30) are referenced in each citing article.

Fortunato, 2011; Wang et al., 2013). Functionally this means that, each time a paper is cited, it is *a priori* more likely to be cited again.

An accurate citation-based measure of influence must address all of these issues, but we reasoned that the key to developing such a metric would be the careful identification of a comparison group, i.e., a cluster of interrelated papers against which the citation performance of an article of interest, or reference article (RA), could be evaluated.

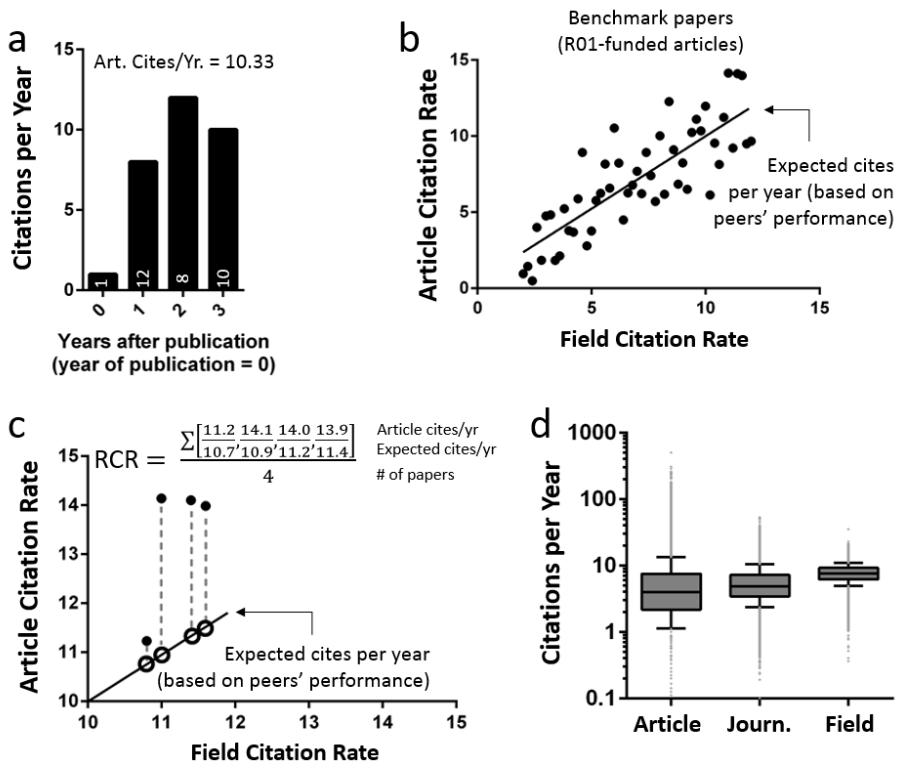


Figure 2. Algorithm for calculating the Relative Citation Ratio. (***a***) Article Citation Rate (ACR) is calculated as the total citations divided by the number of years excluding the calendar year of publication (**Supplemental Equation 1**), when few, if any, citations accrue (**Supplemental Figure 2**). (***b***) Generate an expectation for article citation rates based on a preselected benchmark group, by regressing the ACR of the benchmark papers onto their FCRs (**Supplemental Equations 3, 4**), one regression each publication year. The graphed examples were sampled from a random distribution for illustrative purposes. (***c***) The coefficients from each year's regression equation transforms the Field Citation Rates of papers published in the same year into Expected Citation Rates (**Supplemental Equation 5**). Each paper's RCR is its ACR/ECR ratio. A portfolio's RCR is simply the average of the individual articles' RCRs (**Supplemental Equation 6**). (***d***) Box-and whisker plots of 88,835 NIH-funded papers (published between 2003 and 2010), summarizing their Article Citation Rate, Journal Impact Factor (matched to the article's year of publication), and Field Citation Rate. Boxes show the 25th-75th percentiles with a line at the median; whiskers extend to the 10th and 90th percentiles.

Using a network of papers linked to that RA through citations occurred to us as a promising possibility (**Figure 1**). There are *a priori* three types of article-linked citation networks (Small, 1973). A citing network is the collection of papers citing the RA (**Figure 1a**, top row), a co-citation network is defined as the other papers appearing in the reference lists alongside the RA (**Figure 1a**, middle row), and a cited network is the collection of papers in the reference list of the RA (**Figure 1a**, bottom row).

All three types of networks would be expected to accurately reflect the interdisciplinary nature of modern biomedical research and the expert opinion of publishing scientists, who are themselves the best judges of what constitutes a field. Unlike cited networks, citing and co-citation networks can grow over time, allowing for the dynamic evaluation of an article's influence; they can also indicate whether or not an article gains relevance to additional disciplines (**Figure 1b, c**). An important difference between citing and co-citation networks, however, is size. Papers in the biomedical sciences have a median of 30 articles in their reference list in this dataset, so each citation event can be expected to add multiple papers to an article's co-citation network (**Figure 1d**), but only one to its citing network. The latter are therefore highly vulnerable to finite number effects; in other words, for an article of interest with few citations, small changes in the citing network would have a disproportionate effect on how that article's field was defined. We therefore chose to pursue co-citation networks as a way to describe an individual paper's field.

Calculating the Relative Citation Ratio

Having chosen our comparison group, the next step was to decide how to calculate the values that numerically represent the co-citation network of each RA. The most obvious choice, averaging the citation rates of articles in the co-citation network, would also be highly vulnerable to finite number effects. We therefore chose to average the citation rates of the

journals represented by the collection of articles in each co-citation network. If a journal was represented twice, its journal citation rate (JCR) was added twice when calculating the average JCR. For reasons of algorithmic parsimony we used the JCRs for the year each article in the co-citation network was published; a different choice at this step would be expected to have little if any effect, since almost all JCRs are quite stable over time (**Supplemental Figure 1; Supplemental Table 1**). Since a co-citation network can be reasonably thought to correspond with an RA's area of science, the average of all JCRs in a given network can be redefined as that RA's field citation rate (FCR).

Using this method (**Figure 2a-c; Supplemental Figure 2; Supplemental Equations 1 and 2**), we calculated FCRs for 35,837 papers published in 2009 by NIH grant recipients, specifically those who received R01 awards, the standard mechanism used by NIH to fund investigator-initiated research. We also calculated what the FCR would be if it were instead based on citing or cited networks. It is generally accepted that, whereas practitioners in the same field exhibit at least some variation in citation behavior, much broader variation exists among authors in different fields. The more closely a method of field definition approaches maximal separation of between-field and within-field citation behaviors, the lower its expected variance in citations per year (CPY). FCRs based on co-citation networks exhibited lower variance than those based on cited or citing networks (**Table 1**). Interestingly, a larger analysis of the 88,835 papers published by investigators with continuous R01 funding between 2003 and 2010 shows that FCRs also display less variance than either ACRs ($p < 10^{-4}$, F-test for unequal variance) or JIFs ($p < 10^{-4}$, F-test for unequal variance, **Figure 2d, Table 1**), confirming that co-citation networks are better at defining an article's field than its journal of publication.

Having established the co-citation network as a means of determining an FCR for each RA, our next step was to calculate ACR/FCR ratios. Since both ACR and FCR are measured in CPY, this generates a rateless, timeless metric that can be used to assess the relative influence of any two RAs. However, it does not measure these values against any broader context. For example, if two RAs have ACR/FCR ratios of 0.7 and 2.1, this represents a three-fold difference in influence, but it is unclear which of those values would be closer to the overall mean or median for a large collection of papers. One additional step is therefore needed to adjust the raw ACR/FCR ratios so that, for any given FCR, the mean RCR equals 1.0. Any selected cohort of RAs can be used as a standard for anchoring expectations, i.e. as a customized benchmark (**Supplemental Equations 3-6**). We selected R01-funded papers as our benchmark set; for any given year, regression of the ACR and FCR values of R01-funded papers yields the equation describing, for the FCR of a given RA published in that year, the expected citation rate (**Figure 2b and Supplemental Table 2**). Inserting the ACR as the numerator and FCR of that RA into the regression equation as the denominator is the final step in calculating its RCR value, which incorporates the normalization both to its field of research, and to the citation performance of its peers (**Figure 2b, c and Supplemental Information**).

For analyses where it is important that article RCRs sum to the number of papers for accounting purposes, ordinary least squares (OLS) linear regression of ACR on FCR will benchmark articles such that the mean RCR is equal to 1.0. However, the median article RCR will be lower if there is a skewed distribution. For comparison to the "average" article, quantile regression will yield a median RCR equal to 1.0. OLS regression benchmarking may be more suitable for large-scale analyses conducted by universities or funding agencies, while the quantile regression benchmarking approach might be more suitable for web tools enabling search and exploration at the article or investigator level. In the following analyses, we used OLS regression such that the mean RCR for benchmark articles is equal to 1.0.

Expert validation of RCR as a measure of influence

For the work presented here, we chose as a benchmark the full set of 311,497 RAs published from 2002 through 2012 by NIH-R01 awardees. To measure the degree of correspondence between our method and expert opinion, we compared RCRs generated by benchmarking ACR/FCR values against this standard to three independent sets of post-publication evaluations by subject matter experts (details in **Supplemental Information**). We compared RCR with expert rankings for 2193 articles published in 2009 and evaluated by Faculty of 1000 members (**Figure 3a**), as well as rankings of 430 Howard Hughes Medical Institute- or

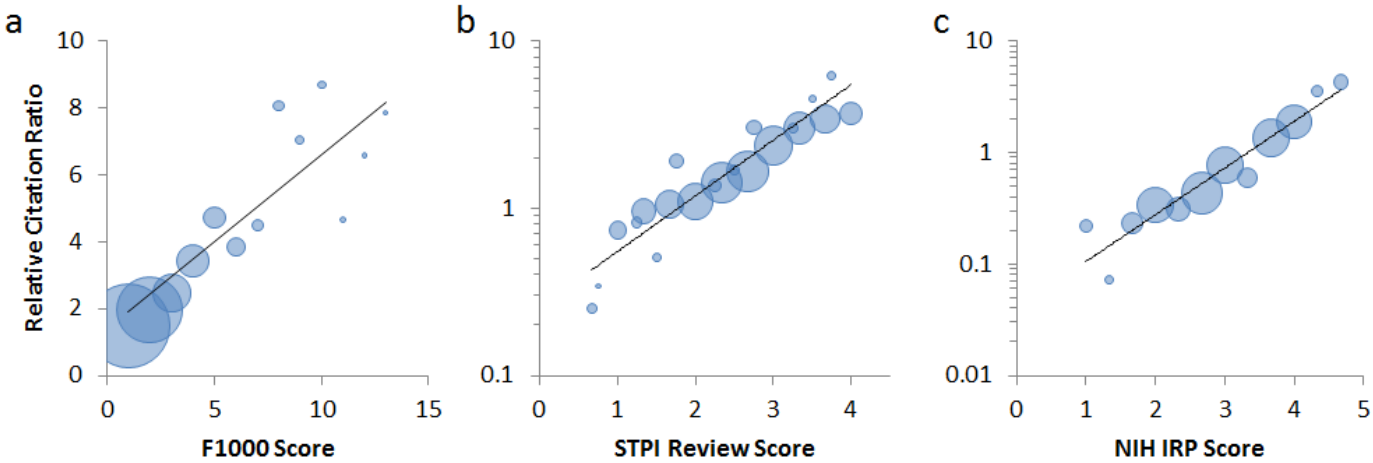


Figure 3. Relative Citation Ratios correspond with expert reviewer scores. (***a-c***) Bubble plots of reviewer scores vs. RCR for three different datasets. Articles are binned by reviewer score; bubble area is proportionate to the number of articles in that bin. (***a***) F1000 scores for 2193 R01-funded papers published in 2009. Faculty reviewers rated the articles on a scale of one to three ("Good", "Very Good", and "Exceptional", respectively); those scores were summed into a composite F1000 score for each article (Supplemental Figure 3). (***b***) Reviewer scores of 430 HHMI and NIH-funded papers collected by the Science and Technology Policy Institute. (***c***) Scores of 290 R01-funded articles reviewed by experts from the NIH Intramural Research Program. Black line, linear regression.

NIH-funded articles published between 2005 and 2011 and evaluated in a study conducted by the Science and Technology Policy Institute (STPI, **Figure 3b**), and finally, 290 articles published in 2009 by extramurally funded NIH investigators and evaluated by NIH intramural investigators in a study of our own design (**Figure 3c; Supplemental Figs. 5-7**). All three approaches demonstrate that RCR values are well correlated with reviewers' judgments. We asked experts in the latter study to provide, in addition to an overall score, scores for several independent sub-criteria: likely impact of the research, importance of the question being addressed, robustness of the study, appropriateness of the methods, and human health relevance. Random forest analysis indicated that their scores for likely impact were weighted most heavily in determining their overall evaluation (**Supplemental Figure 6**).

In addition to correlating with expert opinion, RCR is ranking invariant, which is considered to be a desirable property of bibliometric indicators (Rousseau and Leydesdorff, 2011; Glänzel and Moed, 2012). In short, an indicator is ranking invariant when it is used to place two groups of articles in hierarchical order, and the relative positions in that order do not change when uncited articles are added to each group. The RCR metric is ranking invariant when the same number of uncited articles is added to two groups of equal size (**Supplemental Equations 7-9**). RCR is also ranking invariant when the same proportion of uncited articles is added to two groups of unequal size (**Supplemental Equations 10-11**). This demonstrates that the RCR method can be used effectively and safely in evaluating the relative influence of large groups of publications.

Quantifying how past influence predicts future performance

We next undertook a large case study of all 88,835 articles published by NIH investigators who maintained continuous R01 funding from fiscal year (FY) 2003 through FY2010 to ask how the RCR of publications from individual investigators changed over this eight year interval. Each of these investigators had succeeded at least once in renewing one or more of their projects through the NIH competitive peer review process. In aggregate, the RCR values for these articles are well-matched to a log-normal distribution; in contrast, as noted previously by others, the distribution of impact factors of the journals in which they were published is non-normal (Mansilla et al., 2007; Egghe, 2009) (**Figure 4a, b**). Sorting into quintiles based on JIF demonstrates that, though journals with

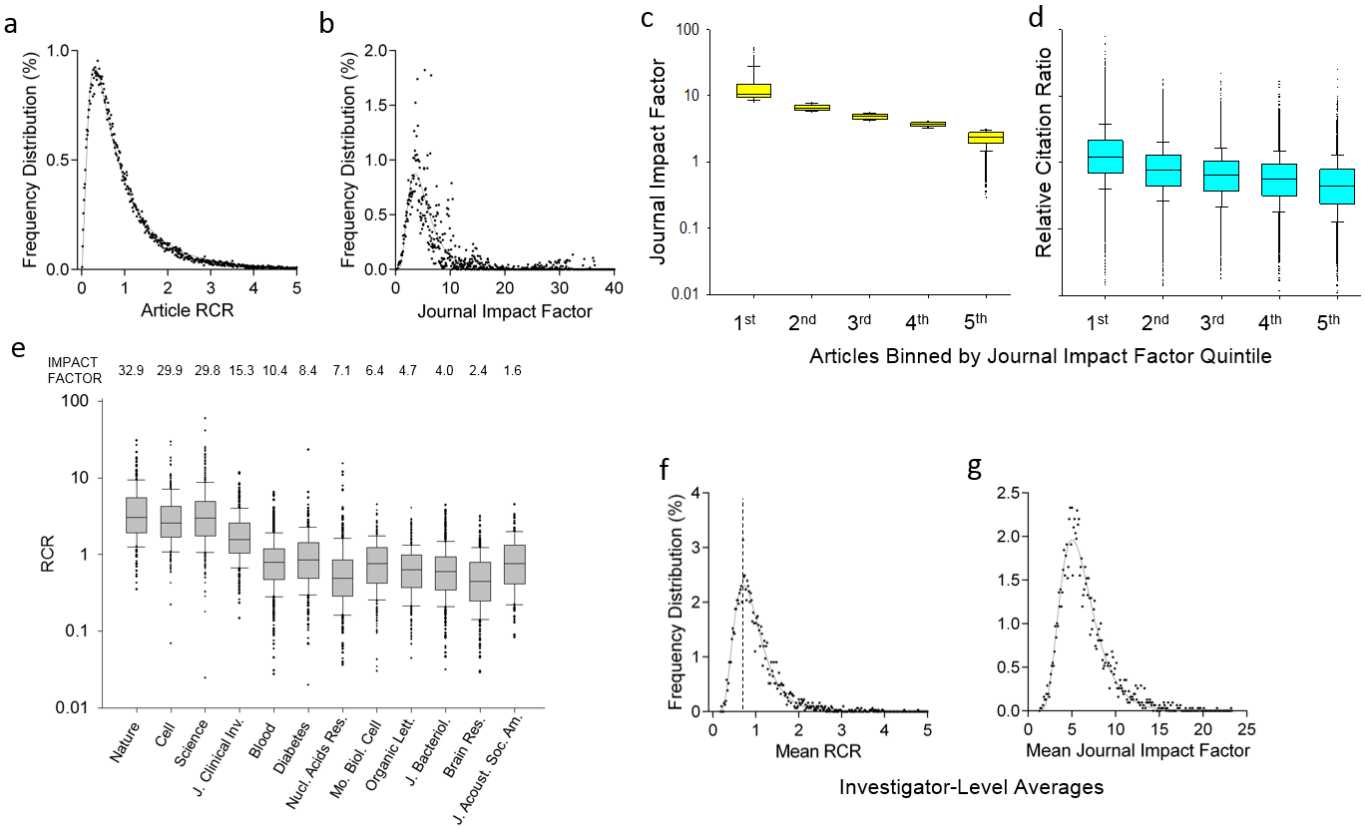


Figure 4. Properties of Relative Citation Ratios at the article and investigator level. (***a, b***) Frequency distribution of article-level RCRs (***a***) and Journal Impact Factors (***b***), from 88,835 papers (authored by 3089 R01-funded PIs) for which co-citation networks were generated. Article RCRs are well-fit by a log-normal distribution ($R^2 = 0.99$), and Journal Impact Factors less so ($R^2 = 0.79$). (***c***) Box-and-whisker plots summarizing Journal Impact Factors for the same papers, binned by Impact Factor quintile (line, median; box, 25th–75th percentiles; whiskers, 10th to 90th percentiles). (***d***) RCR for the same papers using the same bins by Journal Impact Factor quintile (same scale as *c*). Although the median RCR for each bin generally corresponds to the Impact Factor quintile, there is a wide range of article RCRs in each category. (***e***) Box-and-whisker plots summarizing RCRs of these same papers published in selected journals. In each journal, there are papers with article RCRs surpassing the median RCR of the highest Impact Factor journals (left three). The Impact Factor of each journal is shown above. (***f, g***) Frequency distribution of investigator-level RCRs (***f***) and Journal Impact Factors (***g***), representing the mean values for papers authored by each of 3089 R01-funded PIs. Dashed line in (***f***), mode of RCR for PIs.

the highest impact factors have the highest median RCR, influential publications can be found in virtually all journals (**Figure 4c, d**). Focusing on a dozen representative journals with a wide range of JIFs further substantiates the finding that influential science appears in many venues, and reveals noteworthy departures from the correlation between JIF and median RCR (see **Supplemental Information**). For example, NIH-funded articles in both *Organic Letters* (JIF = 4.7) and the *Journal of the Acoustical Society of America* (JIF = 1.6) have a higher median RCR than those in *Nucleic Acids Research* (JIF = 7.1; **Figure 4e**).

As part of this case study we also calculated the average RCR and average JIF for papers published by each of the 3089 NIH R01 principal investigators (PIs) represented in the dataset of 88,835 articles. In aggregate, the average RCR and JIF values for NIH R01 PIs exhibited log-normal distributions (**Figure 4f, g**) with substantially different hierarchical ordering (**Supplemental Figure 8**). This raised a further question concerning PIs with RCR values near the mode of the log-normal distribution (dashed line in **Figure 4f**): as measured by the ability to publish work that influences their respective fields, to what extent does their performance fluctuate? We addressed this question by

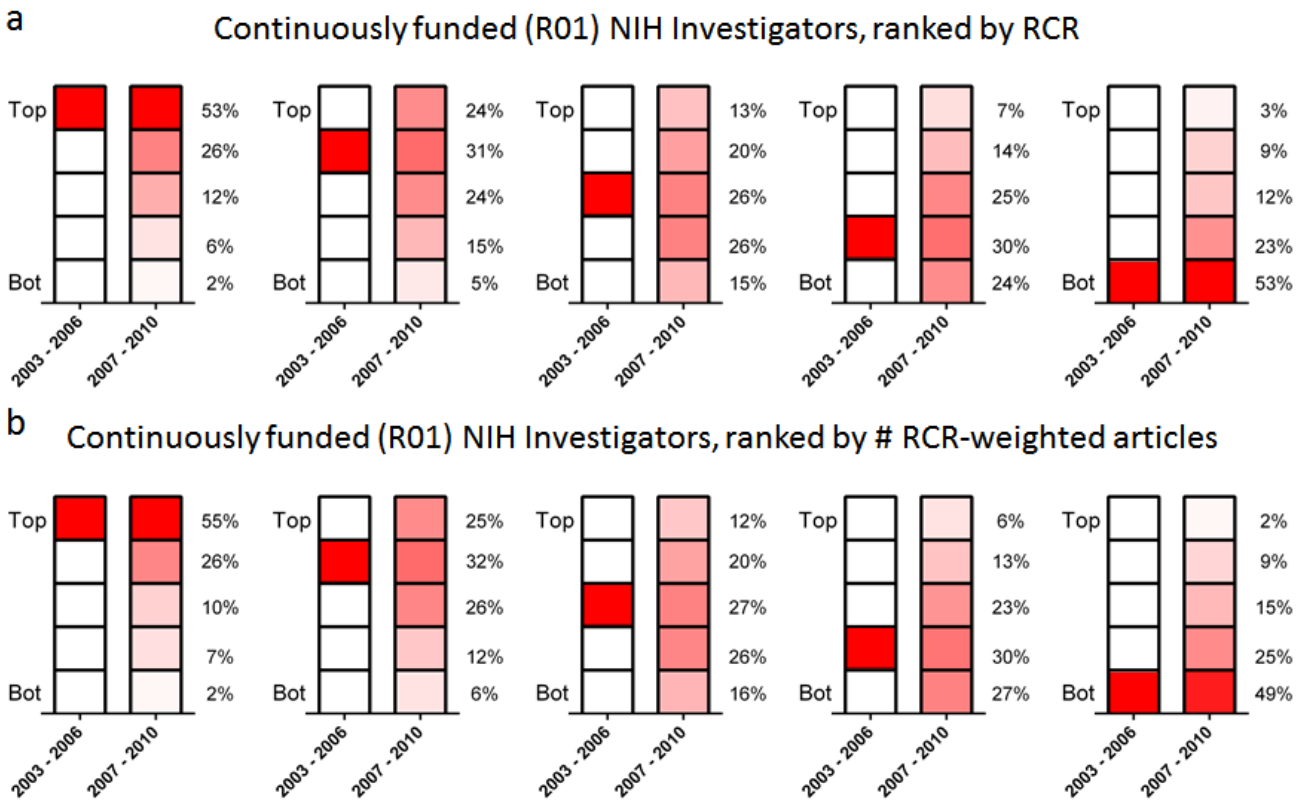


Figure 5. Scientific mobility of investigators' influence relative to their field. Color intensity is proportional to the percentage of PIs in each quintile. (***a***) 3089 investigators who were continuously funded by at least one R01 were ranked by their articles' average RCR in each time window, and split into quintiles. From left to right, investigators starting in different quintiles were tracked to see their rank in the next 4-year period. (***b***) The same analysis, but the number of published articles was multiplied by their average RCR to calculate an influence-weighted article count. PIs were ranked by this aggregate score and split into quintiles.

dividing the eight year window (FY2003 through FY2010) in half. Average RCRs in the first time period (FY2003 through FY2006) were sorted into quintiles, and the percentage of PIs in the second time period (FY2007 through FY2010) that remained in the same quintile, or moved to a higher or lower quintile, was calculated. The position of PIs in these quintiles proved to be relatively immobile; 53% of PIs in the top quintile remained at the top, and 53% of those in the bottom quintile remained at the bottom (**Figure 5a**). For each PI we also calculated a weighted RCR (the number of articles multiplied by their average RCR); comparing on this basis yielded almost identical results (**Figure 5b**). It is worth noting that average FCRs for investigators were extremely stable from one 4-year period to the next (Pearson r = 0.92, **Table 2**), Since FCRs are the quantitative representation of co-citation networks, this further suggests that each co-citation network is successfully capturing the corresponding investigator's field of research.

Discussion

The relationship between scientists and JIFs has been likened to the prisoner's dilemma from game theory: because grant reviewers use JIFs in their evaluations, investigators must continue to weigh this in their decision-making or risk being out-competed by their peers on this basis (Casadevall and Fang, 2014; Shaw, 2014). A groundswell of support for the San Francisco Declaration on Research Assessment (http://www.ascb.org/dora) has not yet been sufficient to break this cycle (Alberts, 2013; Bertuzzi and Drubin, 2013; Schekman and Patterson, 2013;

Suhrbier and Poland, 2013; Casadevall and Fang, 2014; Shaw, 2014). Continued use of the Journal Impact Factor as an evaluation metric will fail to credit researchers for publishing highly influential work. Articles in high-profile journals have average RCRs of approximately 3. However, high-Impact-Factor journals (JIF ≥ 28) only account for 11% of papers that have an RCR of 3 or above. Using Impact Factors to credit influential work means overlooking 89% of similarly influential papers published in less prestigious venues.

Bibliometrics like JIF and H-index are attractive because citations are affirmations of the spread of knowledge amongst publishing scientists, and are important indicators of the influence of a particular set of ideas. These and other prior attempts to describe a normalized citation metric have resulted in imperfect systems for the comparison of diverse scholarly works (Supplemental Information and Supplemental Figure 17), either because they measure only the average performance of a group of papers (Vinkler, 2003), or because the article of interest is measured against a control group that includes widely varying areas of science (Waltman et al., 2011a; Radicchi and Castellano, 2012; Leydesdorff and Bornmann, 2015). An example of the latter is citation percentiling, which the Leiden manifesto (Hicks et al., 2015) recently recommended as best practice in bibliometrics. The RCR method is an improvement over the use of citation percentiling alone, since masking the skewed distribution of citations and article influence, while statistically convenient, can disadvantage portfolios of high-risk, high-reward research that would be expected to have a small proportion of highly influential articles (Rand and Pfeiffer, 2009).

Though tracking the productivity of individual scientists with bibliometrics has been controversial, it is difficult to contradict the assertion that uncited articles (RCR = 0) have little if any influence on their respective fields, or that the best-cited articles (RCR > 20) are impressively influential. We have not determined whether smaller differences, for example those with average or slightly above-average RCRs (e.g. 1.0 versus 1.2), reliably reflect differential levels of influence. Further, citation-based metrics can never fully capture all of the relevant information about an article, such as the underlying value of a study or the importance of making progress in solving the problem being addressed. The RCR metric is also not designed to be an indicator of long-term impact, and citation metrics are not appropriate for applied research that is intended to target a narrow audience of non-academic engineers or clinicians. However, as citation rates mark the breadth and speed of the diffusion of knowledge among publishing scholars, these quantitative metrics can effectively supplement subject matter expertise in the evaluation of research groups seeking to make new discoveries and widely disseminate their findings.

Bibliometric methods also have the potential to track patterns of scientific productivity over time, which may help answer important questions about how science progresses. In particular, co-citation networks can be used to characterize the relationship between scientific topics (including interdisciplinarity), emerging areas, and social interactions. For example, is the membership of an influential group of investigators in a given field or group of fields stable over time, or is it dynamic, and why? Our data demonstrate the existence of an established hierarchy of influence within the exclusive cohort of NIH R01 recipients who remained continuously funded over an eight-year time frame. This may mean that investigators tend to ask and answer questions of similar interest to their fields. Additionally or alternatively, stable differences in investigators' status, such as scientific pedigree, institutional resources, and/or peer networks, may be significant drivers of persistently higher or lower RCR values. Future statistical analyses may therefore reveal parameters that contribute to scholarly influence. To the extent that scientific (im)mobility is a product of uneven opportunities afforded to investigators, there may be practical ways in which funding agencies can make policy changes that increase mobility and seed breakthroughs more widely.

There is increasing interest from the public in the outcomes of research. It is therefore becoming necessary to demonstrate outcomes at all levels of funding bodies' research portfolios, beyond the reporting of success stories that can be quickly and succinctly communicated. For this reason, quantitative metrics are likely to become more prominent in research evaluation, especially in large-scale program and policy evaluations. Questions about how to advance science most effectively within the constraints of limited funding require that we apply scientific

approaches to determine how science is funded (Danthi et al., 2014; Kaltman et al., 2014; Mervis, 2014). Since quantitative analysis will likely play an increasingly prominent role going forward, it is critical that the scientific community accept only approaches and metrics that are demonstrably valid, vetted, and transparent, and insist on their use only in a broader context that includes interpretation by subject matter experts.

Recent work has improved our theoretical understanding of citation dynamics (Radicchi et al., 2008; Stringer et al., 2010; Wang et al., 2013). However, citation counts are not the primary interest of funding bodies, but rather progress in solving scientific challenges. The NIH particularly values work that ultimately culminates in advances to human health, a process that has historically taken decades (Contopoulos-Ioannidis et al., 2008). Here, too, metrics have facilitated quantitation of the diffusion of knowledge from basic research toward human health studies, by examining the *type* rather than the *count* of citing articles (Weber, 2013). Insights into how to accelerate this process will probably come from quantitative analysis. To credit the impact of research that may currently be underappreciated, comprehensive evaluation of funding outputs will need to incorporate metrics that can capture many other outputs, outcomes, and impact, such as the value of innovation, clinical outcomes, new software, patents, and economic activity. As such, the metric described here should not be viewed as a tool to be used as a primary criterion in funding decisions, but as one of several metrics that can provide assistance to decision-makers at funding agencies or in other situations in which quantitation can be used judiciously to supplement, not substitute for, expert opinion.

Materials and Methods

Citation data

The Thomson Reuters Web of Science citation dataset from 2002-2012 was used for all citation analyses. Because of our primary interest in biomedical research, we limited our analysis to those journals in which NIH R01-funded researchers published during this time. For assigning a journal citation rate to a published article, we used the 2-year synchronous journal citation rate (Garfield, 1972; Rousseau and Leydesdorff, 2011) for its journal in the year of its publication. Publications from the final year of our dataset (2012) were not included in analyses because they did not have time to accrue enough citations from which to draw meaningful conclusions, but references from these papers to earlier ones were included in citation counts.

Grant and Principal Investigator data

Grant data was downloaded from the NIH RePORTER database. Grant-to-publication linkages were first derived from the NIH SPIRES database, and the data were cleaned to address false-positives and -negatives. Grant and publication linkages to Principal Investigators were established using Person Profile IDs from the NIH IMPAC-II database. To generate a list of continuously funded investigators, only those Person Profile IDs with active R01 support in each of Fiscal Years 2003-2010 were included.

Calculations and data visualization

Co-citation networks were generated in Python (Python Software Foundation, Beaverton, OR). This was accomplished on a paper-by-paper basis by assembling the list of articles citing the article of interest, and then assembling a list of each paper that those cited. This list of co-cited papers was de-duplicated at this point. Example code for generating co-citation networks and calculating Field Citation Rates is available on GitHub (http://github.com/NIHOPA). Further calculations were handled in R (R Foundation for Statistical Computing, Vienna, Austria). Visualizations were generated in Prism 6 (GraphPad, La Jolla, CA), SigmaPlot (Systat Software, San Jose, CA), or Excel 2010 (Microsoft, Redmond, WA).

When comparing citations rates to other metrics (e.g. post-publication review scores), citation rates were log-transformed due to their highly skewed distribution, unless these other scores were similarly skewed (i.e. Faculty of 1000 review scores). For this process, article RCRs of zero were converted to the first power of 10 lower than the lowest positive number in the dataset (generally 10^{-2}). In the analysis of Principal Investigator RCRs, no investigators had an average RCR of zero.

Table 1. Variance of Field Citation Rates and Expected Citation Rates using different levels of the citation network for calculations (based on 35,837 R01-funded papers published in 2009).

Network level	Variance (FCR)	Variance (ECR)
Impact Factor (RA only)	33.7	30.8
Cited network	16.1	9.5
Citing network	6.7	7.8
Co-citation network	3.4	3.4

Table 2. Summary of investigator-level bibliometric measures and their stability from one 4-year period to the next (PIs with 5 or more articles in each period, except for article count).

Measure	Mean	Median	Pearson *r* (of log-values, '03-'06 vs '07-'10)
Articles per interval (PIs with >0 articles in both 4-yr. intervals)	9.8	8.0	0.56
Field citation rate	7.8	7.7	0.92
Journal citation rate	6.3	5.7	0.76
Article citation rate	6.4	5.3	0.67
RCR	1.0	0.85	0.61

Acknowledgements

We thank Francis Collins, Larry Tabak, Kristine Willis, Mike Lauer, Jon Lorsch, Stefano Bertuzzi, Steve Leicht, Stefan Maas, Riq Parra, Dashun Wang, Patricia Forcinito, Carole Christian, Adam Apostoli, Aviva Litovitz and Paula Fearon for their thoughtful comments on the manuscript, Michael Gottesman for help organizing post-publication peer review, and Jason Palmer, Fai Chan, Rob Harriman, Kirk Baker and Kevin Small for help with data processing and software development for *iCite*.

References

Alberts B (2013) Impact factor distortions. Science 340:787 Available at: http://www.sciencemag.org/content/340/6134/787.full

Anon (2005) Not-so-deep impact. Nature 435:1003–1004 Available at: http://dx.doi.org/10.1038/4351003b

Anon (2013) The maze of impact metrics. Nature 502:271–271 Available at: http://www.nature.com/news/the-maze-of-impact-metrics-1.13952

Bergstrom CT, West JD (2008) Assessing citations with the Eigenfactor metrics. Neurology 71:1850–1851 Available at: http://www.ncbi.nlm.nih.gov/pubmed/19047558

Bergstrom CT, West JD, Wiseman MA (2008) The Eigenfactor metrics. J Neurosci 28:11433–11434 Available at: http://www.ncbi.nlm.nih.gov/pubmed/18987179

Bertuzzi S, Drubin DG (2013) No shortcuts for research assessment. Mol Biol Cell 24:1505–1506 Available at: http://www.molbiolcell.org/content/24/10/1505.long

Bollen J, Van de Sompel H, Hagberg A, Chute R (2009) A principal component analysis of 39 scientific impact measures. PLoS One 4:e6022 Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006022

Bornmann L, Leydesdorff L (2013) The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000. J Informetr 7:286–291 Available at: http://www.sciencedirect.com/science/article/pii/S175115771200106X

Bornmann L, Marx W (2013) How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. Scientometrics 98:487–509 Available at: http://link.springer.com/10.1007/s11192-013-1161-y

Casadevall A, Fang FC (2014) Causes for the persistence of impact factor mania. MBio 5:e00064–14 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3967521&tool=pmcentrez&rendertype=abstract

Contopoulos-Ioannidis DG, Alexiou GA, Gouvias TC, Ioannidis JPA (2008) Medicine. Life cycle of translational research for medical interventions. Science 321:1298–1299 Available at: http://www.sciencemag.org/content/321/5894/1298

Couzin-Frankel J (2013) Shaking up science. Science 339:386–389 Available at: http://www.sciencemag.org/content/339/6118/386.short

Crous CJ (2014) Judge research impact on a local scale. Nature 513:7 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25186869

Danthi N, Wu CO, Shi P, Lauer M (2014) Percentile ranking and citation impact of a large cohort of National Heart, Lung, and Blood Institute-funded cardiovascular R01 grants. Circ Res 114:600–606 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24406983

Egghe L (2009) Mathematical derivation of the impact factor distribution. J Informetr 3:290–295 Available at: http://www.sciencedirect.com/science/article/pii/S175115770900011X

Eom Y-H, Fortunato S (2011) Characterizing and modeling citation dynamics. PLoS One 6:e24926 Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0024926

Garfield E (1972) Citation analysis as a tool in journal evaluation. Science 178:471–479 Available at: http://www.ncbi.nlm.nih.gov/pubmed/5079701

Garfield E (2006) The history and meaning of the journal impact factor. JAMA 295:90–93 Available at: http://jama.jamanetwork.com/article.aspx?articleid=202114

Glänzel W, Moed HF (2012) Opinion paper: thoughts and facts on bibliometric indicators. Scientometrics 96:381–394 Available at: http://link.springer.com/10.1007/s11192-012-0898-z

Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I (2015) Bibliometrics: The Leiden Manifesto for research metrics. Nature 520:429–431 Available at: http://www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci U S A 102:16569–16572 Available at: http://www.pnas.org/content/102/46/16569.full.pdf">JE

Jeong H, Néda Z, Barabási AL (2003) Measuring preferential attachment in evolving networks. Europhys Lett 61:567–572 Available at: http://stacks.iop.org/0295-5075/61/i=4/a=567 [Accessed June 23, 2015].

Johnston M (2013) We have met the enemy, and it is us. Genetics 194:791–792 Available at: http://www.genetics.org/content/194/4/791.long

Kaltman JR, Evans FJ, Danthi NS, Wu CO, DiMichele DM, Lauer MS (2014) Prior publication productivity, grant percentile ranking, and topic-normalized citation impact of NHLBI cardiovascular R01 grants. Circ Res 115:617–624 Available at: http://www.ncbi.nlm.nih.gov/pubmed/25214575

Leydesdorff L, Bornmann L (2015) The operationalization of "fields" as WoS subject categories (WCs) in evaluative bibliometrics: The cases of "library and information science" and "science & technology studies." J Assoc Inf Sci Technol:n/a – n/a Available at: http://doi.wiley.com/10.1002/asi.23408

Mansilla R, Köppen E, Cocho G, Miramontes P (2007) On the behavior of journal impact factor rank-order distribution. J Informetr 1:155–160 Available at: http://www.sciencedirect.com/science/article/pii/S1751157707000296

Mervis J (2014) Peering into peer review. Science 343:596–598 Available at: http://www.sciencemag.org/content/343/6171/596

Misteli T (2013) Eliminating the impact of the Impact Factor. J Cell Biol 201:651–652 Available at: http://jcb.rupress.org/content/201/5/651.short

Moed HF (2010) Measuring contextual citation impact of scientific journals. J Informetr 4:265–277 Available at: http://www.sciencedirect.com/science/article/pii/S1751157710000039

Moed HF, Burger WJM, Frankfort JG, Van Raan AFJ (1985) The use of bibliometric data for the measurement of university research performance. Res Policy 14:131–149 Available at: http://www.sciencedirect.com/science/article/pii/0048733385900125

Opthof T, Leydesdorff L (2010) Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. J Informetr 4:423–430 Available at: http://www.sciencedirect.com/science/article/pii/S1751157710000106

Price DDS (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 27:292–306 Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.161.114

Pulverer B (2013) Impact fact-or fiction? EMBO J 32:1651–1652 Available at: http://emboj.embopress.org/content/32/12/1651.abstract

Radicchi F, Castellano C (2012) Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. J Informetr 6:121–130 Available at: http://www.sciencedirect.com/science/article/pii/S1751157711000794

Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: toward an objective measure of scientific impact. Proc Natl Acad Sci U S A 105:17268–17272 Available at: http://www.pnas.org/content/early/2008/10/30/0806977105.abstract

Rand DG, Pfeiffer T (2009) Systematic differences in impact across publication tracks at PNAS. PLoS One 4:e8092 Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0008092

Rousseau R, Leydesdorff L (2011) Simple arithmetic versus intuitive understanding:The case of the impact factor. ISSI Newsl:10–14 Available at: http://eprints.rclis.org/15478/1/Impact_factor arithmetic.pdf

Schekman R, Patterson M (2013) Reforming research assessment. Elife 2:e00855 Available at: http://elifesciences.org/content/2/e00855.abstract

Seglen PO (1997) Why the impact factor of journals should not be used for evaluating research. BMJ 314:497–497 Available at: http://www.bmj.com/content/314/7079/497.1.short

Shaw D (2014) The prisoners' dilemmas: Authorship guidelines and impact factors: between a rock and a hard place. EMBO Rep 15:635–637 Available at: http://www.ncbi.nlm.nih.gov/pubmed/24781693

Small H (1973) Co-citation in the scientific literature: A new measure of the relationship between two documents. J Am Soc Inf Sci 24:265–269 Available at: http://doi.wiley.com/10.1002/asi.4630240406

Stallings J, Vance E, Yang J, Vannier MW, Liang J, Pang L, Dai L, Ye I, Wang G (2013) Determining scientific impact using a collaboration index. Proc Natl Acad Sci U S A 110:9680–9685 Available at: http://www.pnas.org/content/110/24/9680.short

Stringer MJ, Sales-Pardo M, Nunes Amaral LA (2010) Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. J Am Soc Inf Sci Technol JASIST 61:1377–1385 Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3158611&tool=pmcentrez&rendertype=abstract

Suhrbier A, Poland GA (2013) Are Impact Factors corrupting truth and utility in biomedical research? Vaccine 31:6041–6042 Available at: http://www.sciencedirect.com/science/article/pii/S0264410X13014576

Van Raan AFJ, van Leeuwen TN, Visser MS, van Eck NJ, Waltman L (2010) Rivals for the crown: Reply to Opthof and Leydesdorff. J Informetr 4:431–435 Available at: http://www.sciencedirect.com/science/article/pii/S1751157710000301

Vinkler P (2003) Relations of relative scientometric indicators. Scientometrics 58:687–694 Available at: http://link.springer.com/10.1023/B:SCIE.0000006888.69146.24

Walker D, Xie H, Yan K-K, Maslov S (2007) Ranking scientific publications using a model of network traffic. J Stat Mech Theory Exp 2007:P06010–P06010 Available at: http://www.researchgate.net/publication/2177863_Ranking_Scientific_Publications_Using_a_Simple_Model_of_Network_Traffic

Waltman L, van Eck NJ, van Leeuwen TN, Visser MS, van Raan AFJ (2011a) Towards a new crown indicator: Some theoretical considerations. J Informetr 5:37–47 Available at: http://www.sciencedirect.com/science/article/pii/S1751157710000817

Waltman L, Yan E, van Eck NJ (2011b) A recursive field-normalized bibliometric performance indicator: an application to the field of library and information science. Scientometrics Available at: http://www.akademiai.com/doi/abs/10.1007/s11192-011-0449-z

Wang D, Song C, Barabási A-L (2013) Quantifying long-term scientific impact. Science 342:127–132 Available at: http://www.sciencemag.org/content/342/6154/127.abstract

Weber GM (2013) Identifying translational science within the triangle of biomedicine. J Transl Med 11:126 Available at: http://www.translational-medicine.com/content/11/1/126

Zitt M, Small H (2008) Modifying the journal impact factor by fractional citation weighting: The audience factor. J Am Soc Inf Sci Technol 59:1856–1860 Available at: http://doi.wiley.com/10.1002/asi.20880

Supplement

Supplemental Information 17

Characterization of the co-citation networks of single articles 17

Algorithm and calculations for Relative Citation Ratios (RCRs) 17

Evaluation of different levels of the citation network for field normalization 19

Validation of RCR with post-publication peer review 19

Ranking invariance of RCR 20

Comparison of co-citation vs. journal category denominators 21

Supplemental Figures 22

Supplemental Tables 27

Supplemental References 31

Supplemental Information

Characterization of the co-citation networks of single articles

Field normalization of citations is critical for cross-field comparisons, because of intrinsic differences in citation rates across disciplines. Our approach draws upon the idea of comparing the Article Citation Rate of an article (ACR) with an Expected Citation Rate (ECR), calculated based on peer performance in an article's area of research (Schubert et al., 1986; Vinkler, 2003). Calculating a robust ECR is challenging; other methods frequently employ journals or journal categories as a proxy for a scientific field (Moed et al., 1985; Lundberg, 2007; Zitt and Small, 2008; Opthof and Leydesdorff, 2010; van Raan et al., 2010; Waltman et al., 2011; Bornmann and Leydesdorff, 2013). Unfortunately, these methods do not have sufficient precision to work well at the article level (Leydesdorff and Bornmann, 2015). Other methods have increased granularity of field classification (Waltman and van Eck, 2012; Ruiz-Castillo and Waltman, 2015). However, a granularity threshold must be chosen for analysis, a choice that may differ based on the size of portfolios being compared. Because modern fields of biomedical research exist as a spectrum rather than discrete and separate fields (Talley et al., 2011), we decided to take a more nuanced approach to defining an article's field. We constructed each article's co-citation network (Small, 1973) and used that as a representative sample of its area of research. Simply put, when an article is first cited, the other papers appearing in the reference list along with the article comprise its co-citation network (**Figure 1**). As the article continues to be cited, the papers appearing in the new reference lists alongside it are added to its co-citation network. This network provides a dynamic view of the article's field of research, taking advantage of information provided by the experts who have found the study useful enough to cite.

Algorithm and calculations for Relative Citation Ratios (RCRs)

Our algorithm uses the following steps to calculate RCR values, giving a ratio of ACR to ECR that is benchmarked to papers funded through NIH R01s:

1. Convert the RA citation counts to citations per year (**Figure 2** and **Supplemental Equation 1**).
2. Generate the RA's co-citation network. To do this, we assemble all articles citing the RA; the complete set of papers cited in the reference lists of these citing articles comprises the co-citation network (**Figure 1**).
3. Estimate the FCR of the RA by averaging the journal citation rates of the papers in the co-citation network (**Supplemental Equation 3**).
4. Generate an ECR from the benchmark set of papers. Using R01-funded papers published in a given year, a linear regression of the ACRs vs. FCRs is performed (**Figure 2** and **Supplemental Equations 3-4**). Regressions are calculated for each publication year.
    a. The linear equation coefficients corresponding to the RA's publication year rescale its FCR into a denominator (ECR) that is benchmarked to the performance of R01-funded articles (**Supplemental Equation 5**).
5. The Relative Citation Ratio is the ratio of the ACR : ECR.

When converting raw citation counts to Article Citation Rate (Acr), the year in which the RA was published was excluded from the denominator.

Supplemental Equation 1:

$$Acr = \frac{Total\ citations\ to\ article}{Last\ year\ in\ citation\ database - Year\ of\ article\ publication}$$

We made this design decision because the publication year is nearly always partial, and because articles receive a low number of citations in the calendar year of their publication compared to subsequent years (**Supplemental Figure 2**). In practice, the sum of the citations in years 0 and 1 (the year of publication and the

following year) is close to the mean number of citations per year in the following 8 years (**Supplemental Figure 2**).

ACRs and journal citation rates vary widely from field to field. To compare the Relative Citation Ratios of small groups (like individual investigators), special care must be taken to adjust only the fraction of the citation rate that is due to between-field differences. We tested three methods for adjusting expected citation rates to a field. These approaches each use information from an article's citation network (see schematics in **Figure 1a**). The first method selects the reference article plus those cited in in its reference list (**Figure 1a**, bottom). The second approach instead selects subsequent papers citing the article (**Figure 1a**, top). Finally, the third uses the set of articles that are co-cited with the article of interest by subsequent papers (**Figure 1a**, middle). The Reference Article was always included in the set of papers selected to estimate the FCR, since an article is de facto part of its field. In all three cases, the average of the journal citation rates for the papers in the selected level of the citation network is used as the Field Citation Rate (Fcr).

Supplemental Equation 2:

$$Fcr = \frac{\sum Jcr_i}{N}$$

N is the number of papers in the selected level of the co-citation network (**Figure 1a**) and Jcr_i is the journal citation rate of each paper at the specified level of the citation network.

To generate an expectation of citation performance using this cohort of papers, we performed a linear regression of Article Citation Rate (Acr) in a baseline population against their Field Citation Rates (Fcr) from the same year (**Figure 1**). R01-funded articles were used as a benchmark population. This process was repeated for each year being analyzed to give regression coefficients (slope, $\hat{B}$ and intercept, $\hat{a}$) for benchmarking articles.

Supplemental Equation 3:

$$\hat{B} = \frac{\sum\ (Fcr_i - \overline{Fcr})\ (Acr_i - \overline{Acr})}{\sum\ (Fcr_i - \overline{Fcr})^2}$$

Supplemental Equation 4:

$$\hat{a} = \overline{Acr} - \hat{B}\ \times \overline{Fcr}$$

Alternatively, if benchmarking to the median (rather than mean) field-normalized performance for the benchmark group is desired, quantile regression can be used in lieu of simple linear regression (Koenker, 2005). In either case, the resulting regression line transforms a FCR into an Expected Citation Rate (Ecr^{Year}), corresponding to the ACR that R01-funded papers with the same FCRs and published in the same year were able to achieve, and this can be used outside of the baseline population as a benchmark for articles published in that year:

Supplemental Equation 5:

$$Ecr^{Year} = \hat{B}\ \times Fcr + \hat{a}$$

For the years analyzed here (2002-2011), the resulting regression coefficients are given in Supplemental Table 2. RCR for each article is the ratio of that article's ACR divided by its ECR. Calculating the arithmetic mean is the preferred way of determining the RCR of an entire portfolio (Opthof and Leydesdorff, 2010; van Raan et al., 2010; Waltman et al., 2011):

Supplemental Equation 6:

$$RCR = \frac{1}{n}\sum \frac{Acr_i}{Ecr_i^{Year}}$$

Where n is the number of papers being evaluated, Acr_i is the article citation rate and Ecr_i^{Year} the expected citation rate of each article found by transforming its Fcr_i with the regression coefficients for its publication year (**Supplemental Equation 5**).

Evaluation of different levels of the citation network for field normalization

The aim of this adjustment is to accurately normalize an individual article's citation rate to its field's average citation rate, while preserving within-field differences between papers. Using a set of papers from 2009 matched to R01 grants active in the same year, we compared these three approaches. Calculated RCRs were very similar for all three groups at the article level (**Supplemental Table 3**). This is not surprising, since at this granular scale the article's numerator (ACR) accounts for more of the variance than the denominator. However, accurate field-adjustment is especially important for large-scale analyses, where field differences in citation rates can dominate measurements, as article-level differences in ACR average out. A more accurate estimate of the field citation rate would be predicted to show a smaller correlation between article citation rate and field citation rate, as within-field differences are more effectively excluded. To measure the effectiveness of each level of the citation network, we calculated the correlations between ACRs and ECRs for each approach. Of the three, the "Co-cited" method shows the least correlation between article citations and expected citations (**Supplemental Table 4**). In addition, the variance in expected citation rates should be lower in approaches that more successfully isolate the between-field differences in citation rate from the within-field differences. Again, the co-citation level of the citation network performed the best here (**Table 1**).

Validation of RCR with post-publication peer review

We extensively validated article-level RCRs against expert reviewer scores of the impact or value of papers, using post-publication peer review. Three independently collected sets of post-publication peer reviews were used for this analysis: Faculty of 1000 (F1000) (Li and Thelwall, 2012; Bornmann and Leydesdorff, 2013), a previous survey conducted by the Institute for Defense Analyses Science and Technology Policy Institute (STPI) (Lal et al., 2012), and post-publication peer review conducted by NIH Intramural Research Program (IRP) Principal Investigators.

The first set of expert review scores was compiled from F1000, in which faculty review articles in their fields of expertise, and rate the articles on a scale of 1 to 3 ("Good", "Very Good", and "Exceptional"). Because the decision by the faculty members to review the article is itself a mark of merit, these scores are summed into a composite F1000 score (**Supplemental Figure 3**). We downloaded scores in June 2014 for 2193 R01-funded articles published in 2009 and compared them to their RCRs. This yielded an article-level correlation coefficient ***r*** of 0.44 between RCR and F1000 scores (**Figure 3a**).

For a second set of expert review scores, we took advantage of a previous survey conducted by STPI, of papers funded through the Howard Hughes Medical Institute and NIH. In this survey, experts rated the impact of articles (shown here on a scale of 0 to 4, n = 430 papers from 2005-2011, **Supplemental Figure 4**). Since citation data (including RCR) is highly skewed while survey ratings are not, RCR was log-transformed to bring these ranges into better alignment. The article-level correspondence of RCR with these review scores was similar to that observed with the F1000 scores (***r*** = 0.47, **Figure 3b**).

Finally, we recruited investigators from the NIH Intramural Research Program (IRP) to perform post-publication peer review of NIH-funded articles published in 2009 (**Supplemental Figures 5-7**). A total of 290 articles were independently reviewed by multiple investigators, yielding an article-level correlation of 0.56 (**Figure 3c**). The distribution of these impact scores is shown in Figure 2 Supplement 8. Finally, we asked reviewers in the NIH Intramural Research Program to conduct post-publication peer review of R01-funded articles published in 2009 (reviews conducted by The Scientific Consulting Group, Gaithersburg, MD). Reviewers were asked to give scores on a scale of 1-5 for the following questions:

- Rate whether the question being addressed is important to answer. (1 = Not Important, 2 = Slightly Important, 3 = Important, 4 = Highly Important, 5 = Extremely Important)
- Rate whether you agree that the methods are appropriate and the scope of the experiments adequate. (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree)
- Rate how robust the study is based on the strength of the evidence presented. (1 = Not Robust, 2 = Slightly Robust, 3 = Moderately Robust, 4 = Highly Robust, 5 = Extremely Robust)

- Rate the likelihood that the results could ultimately have a substantial positive impact on human health outcomes. (1 = Very unlikely, 2 = Unlikely, 3 = Foreseeable but uncertain, 4 = Probable, 5 = Almost Certainly)
- Rate the impact that the research is likely to have or has already had. (1 = Minimal Impact, 2 = Some Impact, 3 = Moderate Impact, 4 = High Impact, 5 = Extremely High Impact)
- Provide your overall evaluation of the value and impact of this publication. (1 = minimal or no value, 2 = Moderate value, 3 = Average value, 4 = High value, 5= Extremely high value)

The distribution of responses for each of these questions is shown in **Supplemental Figure 5**. Multiple experts were asked to review each paper, and each set of papers were matched to the fields of expertise of the reviewers examining them. For correlating RCR to review scores, we used the average of the score for the final question (overall evaluation of the paper's value) for articles that were reviewed by at least two experts.

To determine post-hoc which of the first five criteria (importance of scientific question, appropriate methods, robustness of study, likelihood of health outcomes and likely impact) were associated with reviewers' ratings of overall value, we first performed a Random Forest analysis using all 5 criteria. Perhaps unsurprisingly, this analysis showed that "likely impact" was most closely associated with assessment of overall value (**Supplemental Figure 6a**). We subsequently removed this criterion from the analysis to determine the relative importance of the other 4 questions. In this analysis, "Importance" of the scientific question and "Robustness" were most closely linked to overall value (**Supplemental Figure 6b**).

Should an article-level correlation of approximately 0.5 between RCR and expert reviewer scores be considered reliable? This level of correspondence is similar to that previously measured between bibliometric indicators and reviewer scores (Li and Thelwall, 2012; Bornmann and Leydesdorff, 2013). Given the partially overlapping datasets used here, we were able to calculate the correlation of expert reviewer scores with one another. The Pearson correlation coefficient between log-transformed F1000 scores and those from the STPI review was 0.35. In addition, the correlation of scores within a survey can be determined with statistical resampling. We selected papers with three reviews from the STPI and NIH IRP surveys. The order of reviewers was randomly shuffled, and the correlation coefficient between the first reviewer's score and the mean of the other two scores was determined and recorded. This process was repeated 10,000 times for each dataset. The distributions of the 10,000 recorded correlation coefficients are shown in **Supplemental Figure 7**. This approach demonstrated an internal correlation of 0.32 for STPI reviews and 0.44 for the NIH IRP reviews. These values are similar to the correlation between RCR and each set of review scores. These internal correlations between reviewer scores likely represent an estimate of the degree to which it is possible for bibliometrics to correspond to expert opinions. Thus, RCR agrees with expert opinion scores as well as experts agree with one another.

Ranking invariance of RCR

One desirable property in a bibliometric indicator is that of ranking invariance (Rousseau and Leydesdorff, 2011; Waltman et al., 2011; Glänzel and Moed, 2012). A citation metric is ranking invariant if, when two groups of articles are being compared using that indicator, their relative ranking does not change if the groups are inflated by the same amount of uncited papers (Rousseau and Leydesdorff, 2011). RCR is ranking invariant under two cases: when the two comparison groups are the same size and the same absolute number of uncited papers is added to each group, and when the two comparison groups are different sizes and the same proportion of uncited papers is added to each comparison group.

For the first case (two groups of the same size, termed groups I and J, where group I has the greater RCR), the group RCRs are described by the following inequality:

Supplemental Equation 7:

$$\frac{\sum_1^n \frac{Acr_i}{Ecr_i}}{n} > \frac{\sum_1^n \frac{Acr_j}{Ecr_j}}{n}$$

Where i and j are the individual papers from groups I and J, and n is the number of papers in these groups. Adding $k > 0$ papers to each group, each with a constant RCR of $a \geq 0$ (equal to 0 for uncited papers) yields the following inequality:

Supplemental Equation 8:

$$\frac{\left(\sum_1^n \frac{Acr_i}{Ecr_i}\right) + ak}{n+k} > \frac{\left(\sum_1^n \frac{Acr_j}{Ecr_j}\right) + ak}{n+k}$$

This simplifies to **Supplemental Equation 9**, demonstrating ranking invariance under this condition:

Supplemental Equation 9:

$$\sum_1^n \frac{Acr_i}{Ecr_i} > \sum_1^n \frac{Acr_j}{Ecr_j}$$

For the second case (two groups of unequal sizes, termed groups I and J, where group I has the greater RCR), the group RCRs are described by the following inequality:

Supplemental Equation 10:

$$\frac{\sum_1^n \frac{Acr_i}{Ecr_i}}{n} > \frac{\sum_1^n \frac{Acr_j}{Ecr_j}}{m}$$

Where n is the number of papers in group I and m is the number of papers in group J. Adding the same proportion $k > 0$ of papers to each group, each with a constant RCR of $a \geq 0$ (equal to 0 for uncited papers) yields the following inequality:

Supplemental Equation 11:

$$\frac{\left(\sum_1^n \frac{Acr_i}{Ecr_i}\right) + akn}{n(1+k)} > \frac{\left(\sum_1^n \frac{Acr_j}{Ecr_j}\right) + akm}{m(1+k)}$$

This simplifies back to **Supplemental Equation 10**, demonstrating ranking invariance under this condition as well. Note that while uncited papers correspond to $a = 0$, any positive RCR a could be substituted and ranking invariance would hold.

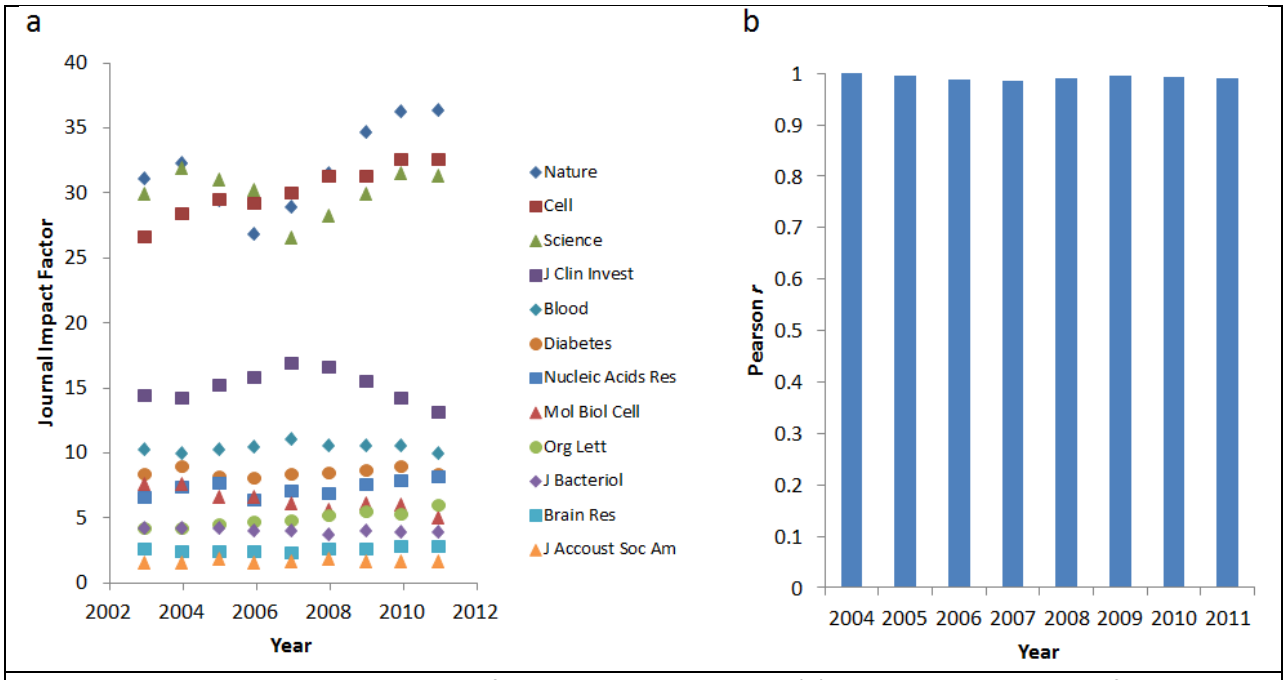
Comparison of co-citation vs. journal category denominators

For generating Relative Citation Ratio-type metrics (a ratio between observed and expected citations), other denominators have been explored in previous work. Early and more recent versions (Schubert et al., 1986; Vinkler, 2003; Neill et al., 2015) used the Journal Impact Factor, or alternatively the average citation count of articles from the same journal, as an expected citation count. Journals tend to be arranged in hierarchies within their fields of research, and journals with higher Impact Factors are generally considered more prestigious. Using articles from the same journal to derive an expected citation boosts the Relative Citation Ratios of those publishing in less prestigious journals, a feature many consider undesirable. Thus, others have segregated journals into categories so that papers from within the entire hierarchy of journals can be used to generate an expected citation count as the denominator (Bornmann and Leydesdorff, 2013; Bornmann and Marx, 2013; Leydesdorff and Bornmann, 2015). In principle, this approach could normalize out differences between fields as well. In practice, journal categorization has been found to be imprecise at the article level (Leydesdorff and Bornmann, 2015), and one reason for this may be that category-based field normalization could go too far in boosting the scores of articles in fields calculated as having a low expected citation rate.

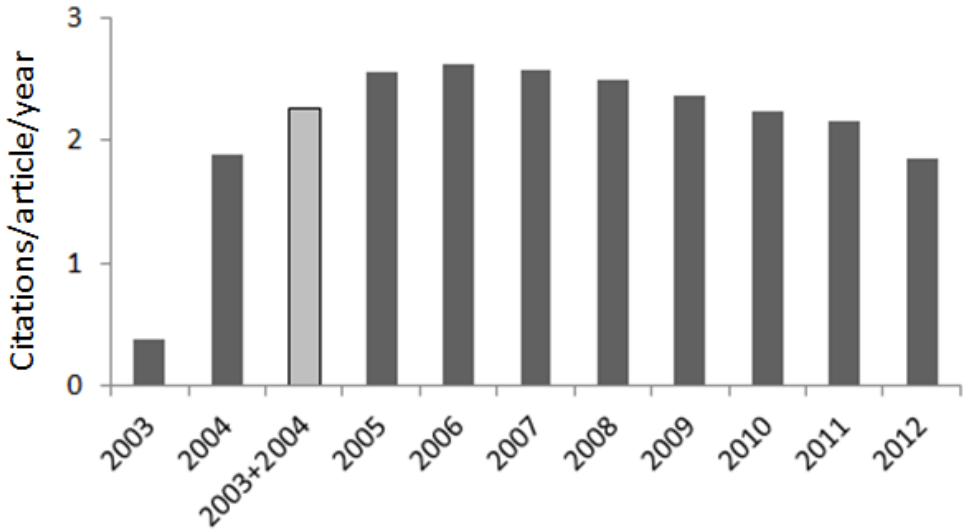
To compare our approach using co-citation networks and regression to R01-funded papers against the journal categorization approach, we examined these two ratios at the low end of field citation rates (< 2.0) in one year of our dataset (2009, with a total of 34,520 R01-funded papers for which we could calculate both metrics (**Supplemental Figure 9**). There were 109 papers with an FCR of less than 2 when calculating RCR, and 544 with a denominator corresponding to an FCR of less than 2 for the journal categorization approach. In general, the journal categorization approach assigned higher scores to these papers (mean score, 1.67, **Supplemental Figure 9**)

compared to RCR (mean RCR, 0.59). The range of values for the journal categorization approach was 0.17 – 23.4, while the range of RCRs was 0.15 – 3.7. These data do not support the conclusion that the Relative Citation Ratio described here overweight articles with low Field Citation Rates compared to previous approaches.

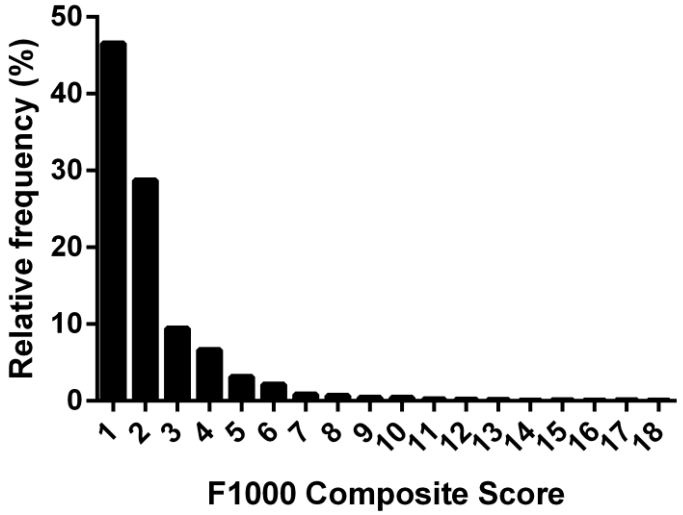
Supplemental Figures



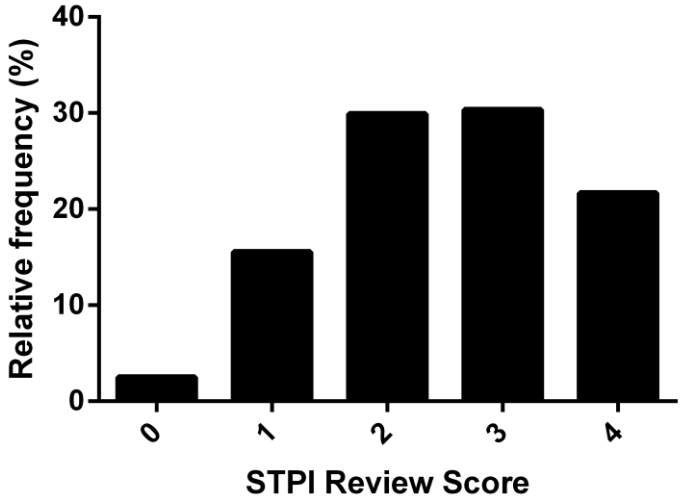
Supplemental Figure 1. Journal impact factor stability over time. (a) Journal Impact Factors for 12 selected journals from 2003 to 2011. (b) Pearson correlation coefficients ***r*** of the Journal Impact Factors for these 12 journals in 2003 vs. each of their respective Impact Factors in subsequent years. In each case, ***r*** is over 0.9.



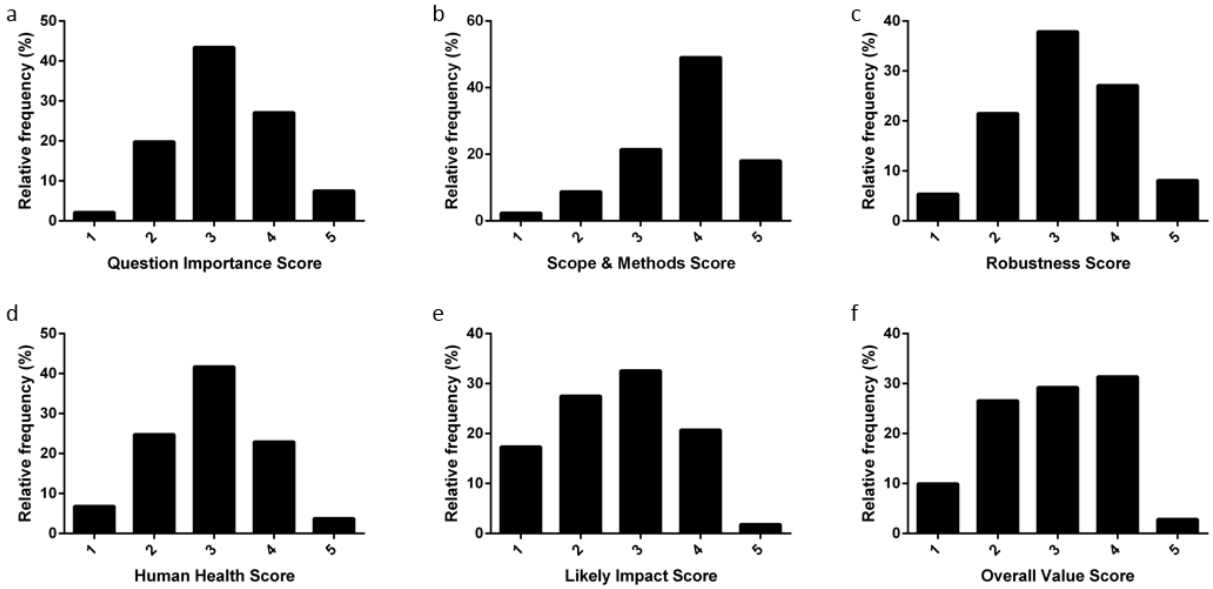
Supplemental Figure 2. Mean citations accrued each year for 608,058 papers published in 2003 appearing in the same journals as NIH-funded publications. Adding the values for 2003 and 2004 gives a value (2.26 citations per publication per year) close to the mean citations per year of the following years (2.36). Although these values may seem low, they are both similar to the global 2013 Aggregate Impact Factor metric for journals appearing in the Biology subcategory (2.56), which is also measured in citations per paper per year.



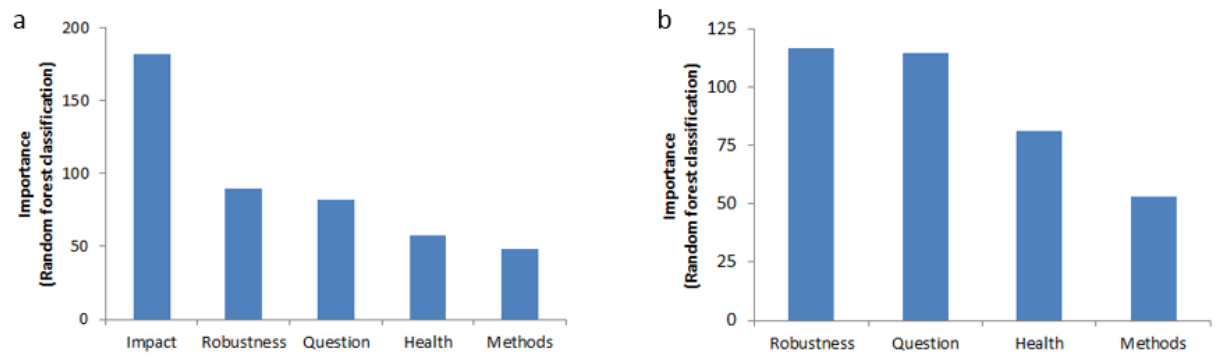
Supplemental Figure 3. Distribution of Faculty of 1000 scores for 2193 R01-funded papers from 2009.



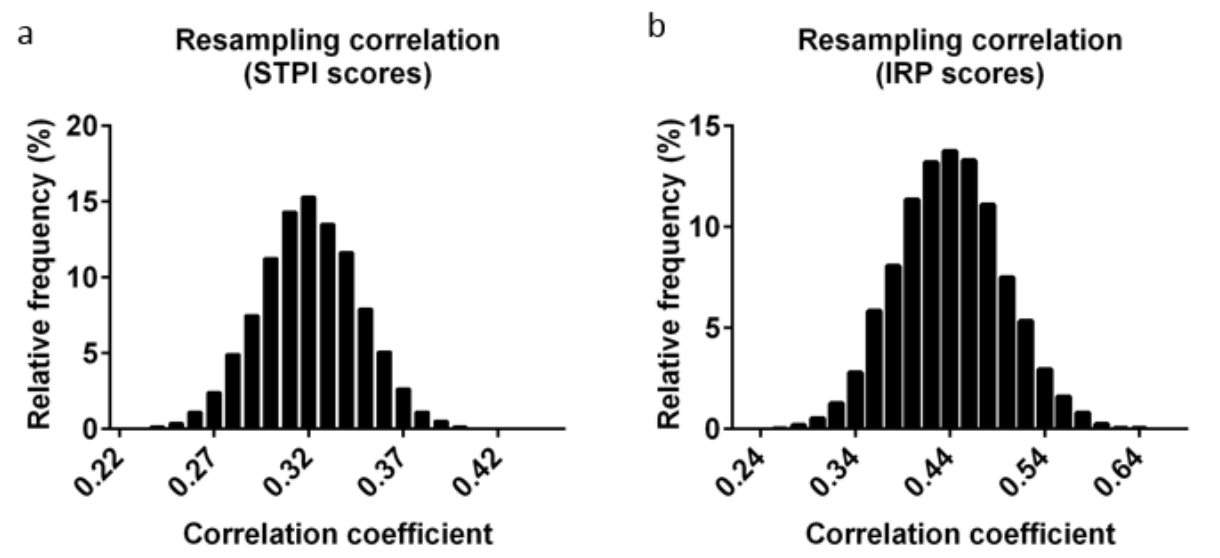
Supplemental Figure 4. Distribution of scores from the Science and Technology Policy Institute survey.



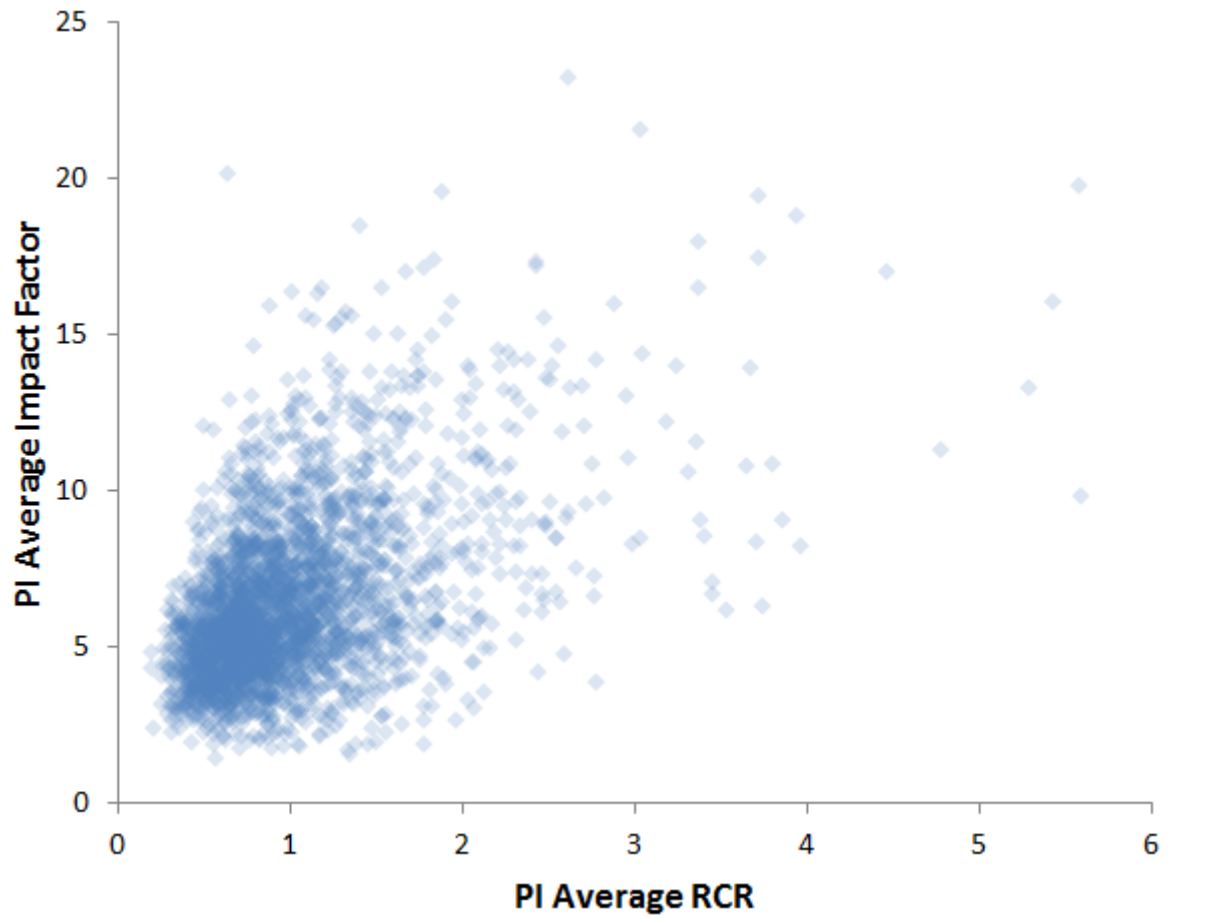
Supplemental Figure 5. Summary of NIH OIRP reviewer responses to post-publication peer review questions. Distribution of ratings to the questions: (a) Rate whether the question being addressed is important to answer. (b) Rate whether you agree that the methods are appropriate and the scope of the experiments adequate. (c) Rate how robust the study is based on the strength of the evidence presented. (d) Rate the likelihood that the results could ultimately have a substantial positive impact on human health outcomes. (e) Rate the impact that the research is likely to have or has already had. (f) Provide your overall evaluation of the value and impact of this publication.



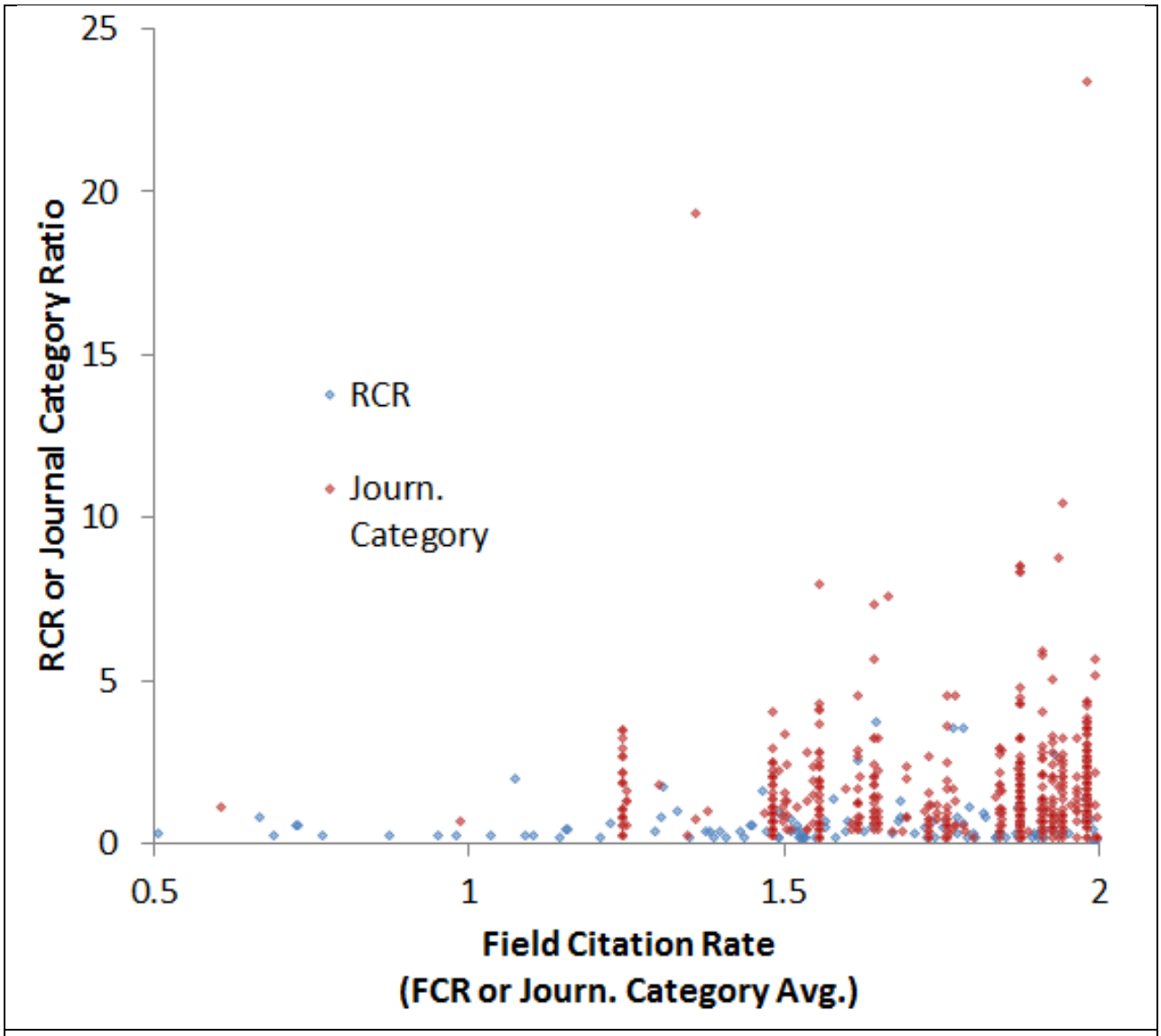
Supplemental Figure 6. Inferred criteria used by NIH IRP reviewers to rate overall value and impact of a publication. (a) Criteria most strongly linked to assessments of "overall value", measured with Random Forest classification. Values indicate the mean decrease in Gini coefficient. (b) Criteria most strongly linked to assessments of "overall value", excluding "likely impact", and measured with Random Forest classification.



Supplemental Figure 7. Internal correlation of post-publication peer review scores is similar to correlation between RCR and review scores. (a) Pearson correlation coefficients (***r***) of one randomly chosen reviewer score vs. the mean of the other two scores for that paper, determined by statistical resampling. Distribution of correlation coefficients determined by resampling from the STPI dataset (10,000 repetitions, mean ***r*** = 0.32). (b) Pearson correlation coefficients (***r***) of one randomly chosen reviewer score vs. the mean of the other two scores for that paper, determined by statistical resampling. Distribution of correlation coefficients determined by resampling from the IRP dataset (10,000 repetitions, mean ***r*** = 0.44).



Supplemental Figure 8. Correlation of the average of different investigators' article RCRs vs. the average of the Journal Impact Factors in which they published. Some investigators published very influential articles (high RCR) in lower-profile venues (low JIF) and vice versa. $R^2 = 0.23$



Supplemental Figure 9. Comparison of RCR and journal categorization approach (Radicchi et al., 2008; Waltman et al., 2011) for deriving expected citations for the ratio denominator. Fewer articles had Field Citation Rates less than 2.0 when calculating RCR (n =109) vs. the journal categorization approach (n = 544), of a possible R01-funded 34,520 papers with the necessary information to calculate both metrics. Data points are partially transparent to allow coordinates with multiple papers (darker) to be more clearly identified.

Supplemental Tables

Supplemental Table 1. Journal Impact Factor stability over time for 100 selected journals.

Journal	2003	2004	2005	2006	2007	2008	2009	2010	2011
ANNU REV IMMUNOL	52.28	52.431	47.4	47.237	47.981	41.059	37.902	49.271	52.761

ANNU REV BIOCHEM	37.647	31.538	33.456	36.525	31.19	30.016	29.875	29.742	34.317
PHYSIOL REV	36.831	33.918	28.721	31.441	29.6	35	37.726	28.417	26.866
NAT REV MOL CELL BIO	35.041	33.17	29.852	31.354	31.921	35.423	42.198	38.65	39.123
NEW ENGL J MED	34.833	38.57	44.016	51.296	52.589	50.017	47.05	53.486	53.298
NATURE	30.979	32.182	29.273	26.681	28.751	31.434	34.48	36.104	36.28
NAT MED	30.55	31.223	28.878	28.588	26.382	27.553	27.136	25.43	22.462
ANNU REV NEUROSCI	30.167	23.143	24.184	28.533	26.077	26.405	24.822	26.756	25.737
SCIENCE	29.781	31.853	30.927	30.028	26.372	28.103	29.747	31.377	31.201
NAT IMMUNOL	28.18	27.586	27.011	27.596	26.218	25.113	26	25.668	26.008
PHARMACOL REV	27.067	22.837	15.689	16.854	18.823	21.936	17	18.861	20.225
CELL	26.626	28.389	29.431	29.194	29.887	31.253	31.152	32.406	32.403
NAT GENET	26.494	24.695	25.797	24.176	25.556	30.259	34.284	36.377	35.532
ANNU REV CELL DEV BI	22.638	17.804	23.69	26.576	23.545	22.731	19.571	14.078	15.836
ANNU REV PHARMACOL	21.786	21.104	19.833	22.808	21.696	21.561	22.468	19.238	21.639
JAMA-J AM MED ASSOC	21.455	24.831	23.494	23.175	25.547	31.718	28.899	30.011	30.026
CHEM REV	21.036	20.233	20.869	26.054	22.757	23.592	35.957	33.036	40.197
NAT CELL BIOL	20.268	20.649	19.717	18.485	17.623	17.774	19.527	19.407	19.488
CANCER CELL	18.913	18.122	18.725	24.077	23.858	24.962	25.288	26.925	26.566
ANNU REV PHYSIOL	18.591	16.672	13.359	15.356	18.875	19.74	18.17	16.106	20.827
LANCET	18.316	21.713	23.878	25.8	28.638	28.409	30.758	33.633	38.278
TRENDS IMMUNOL	18.153	13.075	10.174	10.213	9.48	9.91	8.768	9.533	10.403
NAT REV DRUG DISCOV	17.732	19.583	18.775	20.97	23.308	28.69	29.059	28.712	29.008
NAT BIOTECHNOL	17.721	22.355	22.738	22.672	22.848	22.297	29.495	31.09	23.268
ENDOCR REV	17.324	18.784	22.538	23.901	18.493	18.562	19.761	22.469	19.929
MOL CELL	16.835	16.811	14.971	14.033	13.156	12.903	14.608	14.194	14.178
IMMUNITY	16.016	15.448	15.156	18.306	19.266	20.579	20.589	24.221	21.637
J EXP MED	15.302	14.588	13.965	14.484	15.612	15.463	14.505	14.776	13.853
NAT NEUROSCI	15.141	16.98	15.456	14.805	15.664	14.164	14.345	14.191	15.531
ACCOUNTS CHEM RES	15	13.154	13.141	17.113	16.214	12.176	18.203	21.852	21.64
DEV CELL	14.807	15.434	14.609	13.523	12.436	12.882	13.363	13.946	14.03
NEURON	14.109	14.439	14.304	13.894	13.41	14.17	13.26	14.027	14.736
J NATL CANCER I	13.844	13.856	15.171	15.271	15.678	14.933	14.069	14.697	13.757
TRENDS NEUROSCI	12.631	14.794	14.325	13.494	12.479	12.817	12.794	13.32	14.235
ANN INTERN MED	12.427	13.114	13.254	14.78	15.516	17.457	16.225	16.729	16.733
ANNU REV GENOM HUM G	12.2	8.581	10.094	10.771	10.722	12.029	11.568	17.182	14.829
ANNU REV MICROBIOL	12.105	12.316	13.412	14.553	14.362	10.902	12.804	12.415	14.345
ANNU REV GENET	11.92	11.304	13.959	19.098	18.302	12.78	13.235	21.774	22.233
CLIN MICROBIOL REV	11.53	10.671	10.443	12.643	15.764	16.409	14.691	13.5	16.129
ANNU REV MED	11.381	11.2	10.383	13.237	13.415	10.985	9.94	12.457	13.104
CIRCULATION	11.164	12.563	11.632	10.94	12.755	14.595	14.816	14.432	14.739

J CLIN ONCOL	10.864	9.835	11.81	13.598	15.484	17.157	17.793	18.97	18.372
NAT MATER	10.778	13.531	15.941	19.194	19.782	23.132	29.504	29.92	32.841
ANNU REV PHYS CHEM	10.5	11.944	13.405	11.25	9.439	14.688	17.464	12.245	14.13
ANNU REV PSYCHOL	9.896	12.8	9.784	11.706	13.4	16.217	22.75	18.288	16.833
GENOME RES	9.635	10.382	10.139	10.256	11.224	10.176	11.342	13.588	13.608
PSYCHOL BULL	8.405	7.701	9.746	12.725	10.905	12.568	12.854	11.975	14.457
MOL CELL BIOL	8.142	7.822	7.093	6.773	6.42	5.942	6.057	6.188	5.527
J AM COLL CARDIOL	7.599	9.133	9.2	9.701	11.054	11.438	12.64	14.293	14.156
DIABETES CARE	7.501	7.071	7.844	7.912	7.851	7.349	6.718	7.141	8.087
MOL BIOL CELL	7.454	7.517	6.52	6.562	6.028	5.558	5.979	5.861	4.942
LANCET ONCOL	7.411	8.794	9.608	10.119	12.247	13.283	14.47	17.764	22.589
MOL THER	6.125	5.204	5.443	5.841	5.862	5.97	6.239	7.149	6.873
MOL MICROBIOL	5.563	5.959	6.203	5.634	5.462	5.213	5.361	4.819	5.01
NEUROBIOL AGING	5.552	5.516	5.312	5.599	5.607	5.959	5.937	6.634	6.189
MOL PSYCHIATR	5.539	6.943	9.335	11.804	10.9	12.537	15.049	15.47	13.668
PROG BIOPHYS MOL BIO	5.346	4.551	5.148	5.684	5.009	6.388	3.992	3.964	3.203
THROMB HAEMOSTASIS	4.95	3.413	3.056	2.803	3.501	3.803	4.451	4.701	5.044
J NUCL MED	4.899	5.362	4.684	4.986	5.915	6.662	6.424	7.022	6.381
J BIOL RHYTHM	4.061	2.979	4.367	4.633	3.868	4.211	4.418	3.309	2.934
MOL ECOL	3.87	4.375	4.301	4.825	5.169	5.325	5.96	6.457	5.522
J MAMMARY GLAND BIOL	3.853	2.984	3.625	3.8	3.765	4.167	4.074	5.446	6.741
BIPOLAR DISORD	3.658	4.065	4.812	3.494	4.442	3.959	5.502	5.221	5.289
VITAM HORM	3.439	3.889	4.394	2.24	3.889	3.196	2.439	2.89	2.19
J NEUROSCI RES	3.374	3.727	3.239	3.476	3.268	3.086	2.986	2.958	2.738
AM J PUBLIC HEALTH	3.363	3.241	3.566	3.698	3.612	4.241	4.371	3.85	3.926
DNA REPAIR	3.277	3.92	5.016	5.868	4.018	5.095	4.199	4.293	4.135
MECH DEVELOP	3.254	3.263	3.838	3.836	3.518	2.534	2.827	2.958	2.833
CRIT REV IMMUNOL	3.113	3.595	3.214	3.938	4.058	3.241	2.625	3.857	3.317
LANCET NEUROL	3.07	8.34	11.231	9.479	10.169	14.27	18.126	21.659	23.462
ANTIOXID REDOX SIGN	3.027	3.763	4.232	4.491	5.484	6.19	7.581	8.209	8.456
CELL TISSUE RES	2.991	2.67	2.383	2.58	2.613	2.74	2.308	2.804	3.114
BREAST CANCER RES	2.932	2.975	4.026	4.157	4.371	5.052	5.326	5.785	5.245
BIOMACROMOLECULES	2.824	3.299	3.618	3.664	4.169	4.146	4.502	5.327	5.479
INT J OBESITY	2.794	3.459	4.482	4.055	3.56	3.64	4.343	5.125	4.691
ENDOCRIN METAB CLIN	2.743	2.375	3.685	2.845	2.123	2.121	3.562	4.33	3.411
J CELL BIOCHEM	2.664	2.946	3.591	3.409	3.381	3.54	2.935	3.122	2.868
BRIT J NUTR	2.616	2.71	2.967	2.708	2.339	2.764	3.446	3.072	3.013
BRAIN RES BULL	2.609	2.429	2.481	1.684	1.943	2.281	2.184	2.498	2.818
NEPHROL DIAL TRANSPL	2.607	2.84	2.976	3.154	3.167	3.568	3.306	3.564	3.396
BIOMETALS	2.545	2.155	1.704	1.893	2.17	2.801	3.172	2.32	2.823

INT J ONCOL	2.536	3.056	2.681	2.556	2.295	2.234	2.447	2.571	2.399
STEROIDS	2.444	2.337	2.416	2.849	2.143	2.588	2.905	3.106	2.829
J MED VIROL	2.371	2.331	2.52	2.779	2.831	2.576	2.47	2.895	2.82
NUTR REV	2.318	2.264	2.515	2.937	2.86	3.503	3.443	4.077	4.472
J MEMBRANE BIOL	2.203	2.349	2.208	2.112	2.527	2.32	2.189	1.63	1.808
EUR J CELL BIOL	2.185	2.312	2.195	3.039	3.224	3.955	3.314	3.63	2.806
ANN THORAC SURG	2.041	2.244	2.229	2.342	2.022	2.689	3.644	3.792	3.741
YEAST	1.98	1.941	2.301	1.955	2.619	2.622	1.805	1.626	1.895
NEUROSCI LETT	1.967	2.019	1.898	2.092	2.085	2.2	1.925	2.055	2.105
ARCH VIROL	1.876	1.841	1.819	1.85	1.839	2.02	1.909	2.209	2.111
J VIROL METHODS	1.826	1.729	1.886	2.097	1.933	2.077	2.133	2.139	2.011
J DEV BEHAV PEDIATR	1.699	1.69	1.943	2.17	2.097	2.487	2.265	2.205	2.135
GASTROENTEROL CLIN N	1.684	1.685	1.741	2.019	2.526	2.293	2.558	2.392	2.618
J BIOMAT SCI-POLYM E	1.593	1.255	1.409	1.607	1.862	2.158	2.505	1.842	1.691
CHRONOBIOL INT	1.59	1.521	2.472	2.517	3.771	3.495	3.987	5.576	4.028
LARYNGOSCOPE	1.449	1.576	1.617	1.736	1.801	1.877	2.018	2.096	2.018
J PEDIATR SURG	1.449	1.25	1.125	1.109	1.227	1.557	1.43	1.308	1.45
BRAIN LANG	1.317	1.614	2.129	2.317	2.641	2.929	2.973	3.162	3.115
IN VIVO	0.753	0.811	1.037	1.273	1.143	0.99	1.171	1.159	1.264

Supplemental Table 2. Regression coefficients (ACR on FCR, through 2012), for ordinary least squares linear regression (OLS) or quantile regression to the median (QR) for papers with concurrent R01 funding. One of the criticisms about Impact Factor is that it uses the mean, rather than median, of a skewed distribution (Rossner et al., 2007); for RCR, quantile regression can be used to benchmark articles to the median citation performance in the benchmark group, rather than the mean.

Year	Slope (OLS)	Intercept (OLS)	Slope (QR)	Intercept (QR)
2003	0.497	1.947	0.355	0.870
2004	0.531	1.877	0.344	1.090
2005	0.593	1.645	0.384	0.914
2006	0.608	1.471	0.385	0.967
2007	0.643	1.411	0.381	1.091
2008	0.772	0.711	0.446	0.681
2009	0.768	0.704	0.469	0.542
2010	0.739	0.657	0.490	0.156
2011	0.780	0.120	0.498	0.000

Supplemental Table 3. Correlation coefficients (*r*) between the log-transformed RCR values generated with the three different methods for estimating field citation rates with co-citation networks.

Methods compared	Pearson *r*
Cited ~ Citing	0.928529
Cited ~ Co-cited	0.95274
Citing ~ Co-Cited	0.959203

Supplemental Table 4. Correlation between ACR vs. ECR.

Method used for expected citation rate	**Pearson *r***
Cited	0.292319
Citing	0.250813
Co-cited	0.180248

Supplemental References

Bornmann L, Leydesdorff L (2013) The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000. J Informetr 7:286–291 Available at: http://www.sciencedirect.com/science/article/pii/S175115771200106X

Bornmann L, Marx W (2013) How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. Scientometrics 98:487–509 Available at: http://link.springer.com/10.1007/s11192-013-1161-y

Glänzel W, Moed HF (2012) Opinion paper: thoughts and facts on bibliometric indicators. Scientometrics 96:381–394 Available at: http://link.springer.com/10.1007/s11192-012-0898-z

Koenker R (2005) Quantile Regression. Cambridge University Press. Available at: https://books.google.com/books?hl=en&lr=&id=hdkt7V4NXsgC&pgis=1

Lal B, Wilson AG, Jonas S, Lee EC, Richards AM, Peña VI (2012) An Outcome Evaluation of the National Institutes of Health (NIH) Director's Pioneer Award (NDPA) Program, FY 2004–2006.

Leydesdorff L, Bornmann L (2015) The operationalization of "fields" as WoS subject categories (WCs) in evaluative bibliometrics: The cases of "library and information science" and "science & technology studies." J Assoc Inf Sci Technol:n/a – n/a Available at: http://doi.wiley.com/10.1002/asi.23408

Li X, Thelwall M (2012) F1000, Mendeley and Traditional Bibliometric Indicators. Proc 17th Int Conf Sci Technol Indic Montréal Sci OST 3:541–551 Available at: http://2012.sticonference.org/Proceedings/vol2/Li_F1000_541.pdf

Lundberg J (2007) Lifting the crown—citation z-score. J Informetr 1:145–154 Available at: http://www.sciencedirect.com/science/article/pii/S1751157706000265

Moed HF, Burger WJM, Frankfort JG, Van Raan AFJ (1985) The use of bibliometric data for the measurement of university research performance. Res Policy 14:131–149 Available at: http://www.sciencedirect.com/science/article/pii/0048733385900125

Neill US, Thompson CB, Gibson DS (2015) Assessing Research Productivity. Sci 29:41682 Available at: http://www.the-scientist.com/?articles.view/articleNo/41682/title/Assessing-Research-Productivity

Opthof T, Leydesdorff L (2010) Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. J Informetr 4:423–430 Available at: http://www.sciencedirect.com/science/article/pii/S1751157710000106

Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: toward an objective measure of scientific impact. Proc Natl Acad Sci U S A 105:17268–17272 Available at: http://www.pnas.org/content/early/2008/10/30/0806977105.abstract

Rossner M, Van Epps H, Hill E (2007) Show me the data. J Cell Biol 179:1091–1092 Available at: http://jcb.rupress.org/content/179/6/1091.short

Rousseau R, Leydesdorff L (2011) Simple arithmetic versus intuitive understanding:The case of the impact factor. ISSI Newsl:10–14 Available at: http://eprints.rclis.org/15478/1/Impact_factor arithmetic.pdf

Ruiz-Castillo J, Waltman L (2015) Field-normalized citation impact indicators using algorithmically constructed classification systems of science. J Informetr 9:102–117 Available at: http://www.sciencedirect.com/science/article/pii/S1751157714001126

Schubert A, Glänzel W, Braun T (1986) Relative indicators of publication output and citation impact of european physics research, 1978–1980. Czechoslov J Phys 36:126–129 Available at: http://link.springer.com/10.1007/BF01599744

Small H (1973) Co-citation in the scientific literature: A new measure of the relationship between two documents. J Am Soc Inf Sci 24:265–269 Available at: http://doi.wiley.com/10.1002/asi.4630240406

Talley EM, Newman D, Mimno D, Herr BW, Wallach HM, Burns GAPC, Leenders AGM, McCallum A (2011) Database of NIH grants using machine-learned categories and graphical clustering. Nat Methods 8:443–444 Available at: http://dx.doi.org/10.1038/nmeth.1619

Van Raan AFJ, van Leeuwen TN, Visser MS, van Eck NJ, Waltman L (2010) Rivals for the crown: Reply to Opthof and Leydesdorff. J Informetr 4:431–435 Available at: http://www.sciencedirect.com/science/article/pii/S1751157710000301

Vinkler P (2003) Relations of relative scientometric indicators. Scientometrics 58:687–694 Available at: http://link.springer.com/10.1023/B:SCIE.0000006888.69146.24

Waltman L, van Eck NJ (2012) A new methodology for constructing a publication-level classification system of science. J Am Soc Inf Sci Technol 63:2378–2392 Available at: http://doi.wiley.com/10.1002/asi.22748

Waltman L, van Eck NJ, van Leeuwen TN, Visser MS, van Raan AFJ (2011) Towards a new crown indicator: Some theoretical considerations. J Informetr 5:37–47 Available at: http://www.sciencedirect.com/science/article/pii/S1751157710000817

Zitt M, Small H (2008) Modifying the journal impact factor by fractional citation weighting: The audience factor. J Am Soc Inf Sci Technol 59:1856–1860 Available at: http://doi.wiley.com/10.1002/asi.20880