Scientist Selection – the Process and Impact of Productivity Measures at DBNBR: Does NIH Peer Review Predict Citation Impact?

By

Jonathan D. Pollock, Ph.D. *
Genetics and Molecular Neurobiology Research Branch
Division of Basic Neuroscience and Behavioral Research
National Institute on Drug Abuse
6001 Executive Blvd, Rm 4255
Bethesda, MD 20892 (For FEDEX Rockville, MD 20852)
tel. 301-435-1309
fax. 301-594-6043email.
jpollock@mail.nih.gov

Jeng-Jong Pan, Ph.D.
Clinical Trials Network
National Institute on Drug Abuse
6001 Executive Blvd
Bethesda, MD 20892
(Current Address: Veterans Health Administration,
Washington, DC 20024

Prasad Kothari, M.S.
Synergy Enterprises, Inc.
8757 Georgia Avenue
Suite 1440
Silver Spring, MD 20910

Ananth Charya, M.S.
Office of the Director
National Institute on Drug Abuse
6001 Executive Blvd
Bethesda, MD 20892

Acknowledgement. The authors wish to thank Susan Volman, Da-Yu Wu for their assistance in collecting the data from the Web of Science data and Paul Wakim for his helpful discussion on statisical analysis. The authors also wish to thank Deborah Lewis for providing preliminary scores of the Molecular Neuropharmacology and Signaling Study Section for the June and October 2009 review dates. We would also like to thank Regina Fitzpatrick and Henry Small at Thomson Reuters for their assistance with the Access database.

Abstract

A central question in evaluating peer review is whether peer review is able to predict which proposal and investigator will produce the best science, have the greatest impact, and be most productive. Grant applications submitted to the Division of Basic Neuroscience and Behavioral Research (DBNBR) at the National Institute on Drug Abuse, NIH in FY 2006 and applications reviewed by the Molecular Neuropharmacology and Signaling Study Section for October 2009 and January 2009 council were analyzed to determine whether peer review score predicted citation impact as measured by total number of publications, the total number of citations, the number of citations per publication, average citations per year, h-index, c-index, and average percentile. This analysis showed that the total number of publications, and the total number of citations are highly correlated with each other but not with citations per publication, the c-index, and average percentile. Citations per publication, the c-index and average percentile were more strongly related to each other. This suggests that at least two factors are being measured by the H-index and citations per publication. The total number of publications, total number of citations, and H-index for a five year period time are good predictors of these variables for the next five year period of time while citations per publication, average percentile, and c-index is not. The H-index for 1996-2000 explains 42% of the variance of the H-index for 2001-2005; the total number of citations for 1996-2000 explains 55% of the variance in the number of citations for 2001-2005; and the total number of publications in 1996-2000 explains 64% of the variance in the total number of publications for the next five years. The average percentile, citations per publication, and c-index for a 5 year period weakly predicts future performance, explaining 9%, 7%, and 3% of the variance in the same statistic for the next five years, respectively. The H-index is further validated by the observation that HHMI neuroscience investigators, many of whom are members of the National Academy of Sciences and Noble laureate had significantly higher H-index scores than DBNBR applicants and 2006/2007 Pioneer awardees even when normalizing for years since first publication. The 01-05 citation impact data did not predict score or percentile; less than 2% of the variance in score or percentile explained by any of the 01-05 citation impact statistics. Although no association was observed between score and citation impact data or between percentile and citation data, funded investigators had on average higher citation impact scores than unfunded investigators with the exception of the c-index. Furthermore, the H-index predicted the probability of receiving the award. Partial correlation and multiple regression analysis of R01 applications reviewed by the Molecular Neuropharmacology and Signaling Study Section for October 2009 and January 2009 council suggests that peer review weights scores most heavily on approach and innovation. In this data set correlation of peer review score was somewhat greater than in FY2006 data set for 01-05 but still modest. Inter-rater reliability for the criterion score was also modest with an average standard deviation for average preliminary score 1.17 for each application. Recommendations are made to weight and combine bibliographic measures with peer review scores.

Introduction
In 2007 the National Institutes of Health conducted a review of its peer review system. Input from NIH staff and stakeholders led to the release of set of recommendation for enhancing peer review in March 2008. Based on these recommendations NIH articulated four goals (http://enhancing-peer-review.nih.gov):

The first goal is to engage the best reviewers. To achieve this goal to retain and recruit the best reviewers NIH has given the reviewers greater time flexibility in serving their tenure as reviewers on study sections. Reviewers have the option of serving 12 times on study sections over a 6 year term instead of four. In some cases travel times and burden have been reduced by permitting asynchronous electronic reviews and video enhanced discussion. Best practices for recruiting reviewers and standardizing the training of reviewers have been implemented.

The second goal is to improve the quality and transparency of review. To attain this goal NIH changed the scoring system from a 5 point to 9 point scale that gives a final score as well as scores for each criteria with all applications being scored even if not discussed. A structured critique template was also instituted to provide written feedback to the applicants about the strengths and weaknesses of the application. The length of the application was also shortened and the number of revised applications reduced to one from two. Applications lacking significant merit or have severe ethical concerns may be designated NRFC (not recommended for further consideration).

The third goal is to ensure balanced & fair reviews across scientific fields & career stages, reduce administrative burden, and encourage innovative and original research. To reach this goal NIH has reduced the number of resubmissions to one with the hope reviewers will decide sooner which applications are most meritorious. To increase the number of young investigators NIH has established the category of early stage investigators who are within ten years of their terminal degree and is implementing goals to fund these investigators. Furthermore, applications from early stage investigators as well as for clinical research will be clustered together when possible.

The fourth goal is continuous review of peer review. To implement this goal NIH plans “to conduct on line surveys of stake holders, develop new metrics to track key elements of peer review changes, create data driven mechanisms to evaluate review, and develop peer review pilots and evaluate these pilots.

A central question in evaluating peer review is whether peer review is able to predict which proposal and investigator will produce the best science, have the greatest impact, and be most productive. One possible method to evaluate peer review is to use bibliographic and citation data. Publications that are cited frequently are likely to be those that have large utility and scientific impact. As such, citation impact is an independent form of peer review and should be correlated with the peer review evaluation.

Bibliographic databases such a Scopus and Web of Science now make it possible to obtain citation and publication data of an investigator such as the total number of publications, the total number of citations, the number of citations per publication, average citations per year, h-index, the c-index, and percentile. The h-index develop by (Hirsch, 2005) is defined as the number of publications that have at least h citations. As described by Hirsch publications are ranked from highest to lowest according to the number of citations. By examining where the rank equals the number of citations H is ascertained. Thus, an h of 5 means that the top 5 cited papers are cited each at least 5 times.

These metrics are superior to that of using journal impact factors for evaluating an individual investigator (Jeang, 2007;Jeang, 2009;Seglen, 1997;Garfield, 2006;Jeang, 2007;Jeang, 2009;Seglen, 1997;Pendlebury, 2009). Publishing in a high impact journal does not necessarily imply high impact by a paper appearing in that journal. Journal citation impact scores are heavily influenced by highly cited papers at the upper end of the citation distribution and many papers are not cited (Pendlebury, 2009).

In this report we analyze the citation impact of investigators who submitted investigator initiated R01 applications to the Division of Basic Neuroscience and Behavioral Research at the National Institute on Drug Abuse, NIH in FY 2006. We analyzed how the total number of publications, the total number of citations, the number of citations per publication, average citations per year, h-index, the c-index, and percentile are related to each other. We then tested whether these citations statistic correlate with peer review and determined whether citation impact statistics have validity as measuring the quality of investigators. Last, we determine which impact statistics predict future performance.

Method

Analysis was conducted on 470 R01 applications to the Division of Basic Neuroscience and Behavioral Research (DBNBR) for funding in FY 2006. These investigator initiated R01 applications evaluated by peer review were given a priority score from 100 to 500 and a percentile by Initial Review Groups (IRG). The priority score of 100 is the best score and the 1st percentile represents the best rank. Applications that are within the bottom half are not discussed or scored. Data for the application number, the name of the PI, the status as an amended application, priority score and percentile, and award status were obtained by searching QVR in ImpacII, a NIH database.

To obtain data for citation analysis the number of publications, the number of citations, the average number of citations per publication, H-index for the investigator for all publications up to 2006, the number of years publishing , the average citations per year were obtained by searching ISI Web of Science. Citation data for each investigator was manually obtained by searching on the last name and the initials of the investigator together with institutional affiliations. Institutional affiliations for each investigator were obtained from the biosketches contained in the grant application. All citation data is based on articles, reviews, editorial materials, notes, and abstracts curated by the Web of Science.

Citation impact data from the 62 neuroscientists supported by the Howard Hughes Medical Institute and 25 Pioneer Awardees in 2006 and 2007 was obtained from the ISI Web of Science by the same method used to curate citation impact data from the 469 DBNBR applications.

Five year citations of papers published from 1996 to 2000 and from 2000 to 2005 were obtained in which citations were restricted to those 5 years. Thus, citation data to papers published from 1996 to 2000 or from 2000 to 2005 only includes citations by papers published in that 5 year period and not later. Because this data is not easily obtained from the public version of the Web of Science, last names of the investigators and initials together with the institutional affiliations were provided to Thomson Reuters, the publishers of the Web of Science. Thompson Reuter under contract to NIH returned the data as Access database for 371 investigators. From the data provided the H-index for citation within the 5 year period, the average citations per publication for that 5 year period, the citations per year, percentile ranking for that 5 year period, and c-index for that 5 year period was obtained for each investigator. The percentile ranking is based on citation ranking for a group of investigator that published in a defined set of journals. The c-index is the number of citations for a given paper subtracted from the expected number of citations for that journal. Initial citation analysis is based on all articles, reviews, editorial materials, notes, and abstracts.

All data were placed in Excel spreadsheets and the statistical programs in Excel were used to compute descriptive statistics. No inferential statistics were used because of the large number of comparisons being made.

To determine how reviewers weight significance, investigator, approach, innovation, and environment influence in determining their preliminary scores, correlation, partial correlation, and regression analysis were performed The sample examined was comprised of 71 investigators who submitted 73 investigator-initiated R01 research grant applications reviewed by the Molecular Neuropharmacology and Signaling study section in June and October of 2009. The top half of applications were assigned a priority score from 10 to 90 and a corresponding percentile, while the bottom half of the applications were not discussed or scored. Each reviewer for an application provides preliminary scores for five evaluative criteria: 1) Significance, 2) Investigator, 3) Innovation, 4) Approach, and 5) Environment, which contribute to the determination of their initial overall score. The scores for the above 5 categories and the initial overall score range from 1 to 9. Data for the application number, the name of the PI, the status as an amended application, priority score and percentile, council date and award status were obtained by searching QVR in ImpacII, a NIH database.

Data for bibliographic measures for each investigator, including h-index, h-index over 2004-2008, and citations per publication were obtained by searching on investigator last name and initials on ISI Web of Knowledge. Where necessary, results were narrowed by including investigator institution as a search term. In addition, to ensure a more accurate identification of each investigator’s publication record, the “distinct author set” option was used, which attempts to remove ambiguity by combining factors such as author name, field of study, and length of publication career to create clusters of articles likely written by an author. Institutional affiliations for each investigator were obtained from the biosketches contained in the grant application. All citation data are based on articles, reviews, editorial materials, notes, and abstracts curated by the Web of Science.

For the purposes of the correlation and multiple regression analysis, records with incomplete data, including missing criterion or priority scores, were removed (n = 2 and n=2, respectively). All statistical analyses were performed with SAS 9.1 (SAS Institute, Cary, NC).

Results

Sample characteristics

The applications analyze the field of basic neuroscience and behavioral research with a focus on drug abuse and addiction. Table 1 displays the application status of the 470 applications submitted in FY2006. 90 of the 470 applications submitted were awarded in FY2006 or (19%). 81 of the 470 were approved by the IRG, 178 of the 470 applications were unscored, and another 121 applications were withdrawn. 81 of the 121 withdrawn applications were first submission of new and competing renewal of R01 applications with the remaining 40 applications withdrawn as first amended applications. The total number of applications scored for the first submission, for the first amended application, and the second amended application is 137, 110, and 45, respectively. Of the 178 unscored applications, 116 were first time submissions of new and competing R01s; 51 were first amended applications, and 11 were second amended applications.

	Approved by IRG	Awarded	Unscored	Withdrawn
Type 1-01	24	14	109	50
Type 1-02	20	24	40	21
Type 1-03	13	15	10	0
Type 2-01	7	11	7	31
Type 2-02	5	21	11	19
Type 2-03	12	5	1	0
Total	**81**	**90**	**178**	**121**

Table 1. Application status of R01 application submitted to the Division of Basic Neuroscience and Behavioral Research at the National Institute on Drug Abuse for consideration for funding.

For purposes of doing correlation analysis for citation impact and its correlation with peer review, investigators that had submitted either the same application or more than one application were removed, leaving a sample size of 362 applications. The total number of publications, total number of citations, citations per publication, H-index, number of years since first publication, publications per year, the number of papers published from 1996 to 2000, the number of paper published in 2001 to 2005, average citations per year, and m=H-index/years since fist publication are shown in Table 2.

Correlation (N=362)	Total Pub	Total Cit	Avg Cit/Pub	H-index	Years publishing	avg pubs/year	Num pubs '96-'00	Num pubs '01-'05	Cit/Year	m=H/Years
Total Pub	1.00									
Total Cit	0.80	1.00								
Avg Cit/Pub	0.09	0.42	1.00							
H-index	0.83	0.91	0.40	1.00						
Years publishing	0.58	0.40	0.02	0.57	1.00					
avg pubs/year	0.90	0.76	0.10	0.78	0.31	1.00				
Num pubs '96-'00	0.89	0.75	0.13	0.79	0.47	0.86	1.00			
Num pubs '01-'05	0.80	0.62	0.02	0.68	0.41	0.86	0.82	1.00		
Cit/Year	0.67	0.95	0.55	0.86	0.23	0.73	0.68	0.59	1.00	
m=H/Years	0.42	0.63	0.53	0.66	-0.13	0.65	0.48	0.46	0.78	1.00

Table2: Correlation between data spanning all years to 2006 after cleaning (N=362)

The total number of publications for all years produced by an investigator is strongly associated with the number of citations (r^2=0.64), the number of publications per year (r^2=0.81), the H-index (r^2=0.69), number of publications 1996-2000 (r^2=0.79), the number of publications 2001-2005 (r^2=0.64) but moderately associated with the number of years publishing (r^2=0.34) and citations per year (r^2=0.45). The total number of publications is weakly correlated with both m (r^2=0.18) and the average number of citations per publication (r^2=0.01).

The total number of citations produced by an investigator is strongly correlated with number of publications (r^2=0.64), number of publications 1996-2000 (r^2=0.56), H-index (r^2=0.83), citations per year (r^2=0.90), the average number of publications per year (r^2=0.58) but modestly with number of publications 2001-2005 (r^2=0.38), and m (r^2=0.40). The average number of citations per publication and the number of years publishing are weakly associated with the total number of citations, (r^2=0.18) and (r^2=0.16), respectively.

The average number of citations per publication is at best modestly associated with any other citation measure. The best correlations are obtained between citations per publication and average citations per year is (r^2=0.30) and m (r^2=0.28). The association between average citation per publication and total citations is (r^2=0.18) or with the H-index is (r^2=0.16). The amount of the variance for citation per publications explained by number of years publishing is (r^2=0.00) while the amount of the variance for citation per publication explained by average publications per year is (r^2=0.01). The fraction of the variance in the number of citations per publication explained by the number of publications 1996-2000 or the number of publications 2001-2005 is very small, 2 and 0 percent, respectively.

The average number of citations per year is correlated with other measures of citation impact. Total number of publications, total number of citations, citations per publication, H-index, years publishing, and average number of publications per year explains 44%, 90%, 30% 74%, 5%, and 53% of the variance in average citations per year, respectively. The average number of citations per year is moderately associated with number of publications from 1996-2000 and 2001-2006 r^2=0.46 and r^2=0.35, respectively.

The average number of publications per year for all years is strongly correlated with either the total number of publications (r^2=0.81), with the number of publications from 1996 to 2000 (r^2=0.74), with the number of publications from 2001 to 2005 (r^2=0.74) and with the total citations (r^2=0.58), with citations per year (r^2=0.53), with m (r^2=0.42), or with the H-index (r^2=0.62). The average number of publications is weakly associated with either citations per publication (r^2=0.01) and the number of years publishing (r^2=0.10).

The H-index is defined as the number of papers published by an investigator that are cited at least h times. Thus, an H of 10 means that ten papers are cited at least ten times and there are a least a total of 100 citations. The H-index for investigators submitting applications for all years was highly correlated with the total number of publications (r^2=0.69), with the number of publications from 1996 to 2000 (r^2=0.62), with the total number of publications from 2001 to 2005 (r^2=0.42), with the total number of citations (r^2=0.83), with the average number of publications per year (r^2=0.61), with the number of citations per year (r^2=0.74). There is also a moderate positive correlation between an investigator's H-index and the number of years since the investigator published his first paper (r^2=0.32). The H-index is weakly associated with the average number of citations per publication (r^2=0.16). The correlation coefficients for the H-index with total number of publications, total number of citations, citations per year, the number of year publishing, and publications per year are remarkably similar to those obtained for HHMI investigators and Pioneer awardees (Table 3). The high correlations and the reproducibility of the correlations in different samples suggest that H-Index will produce reliable results regardless of the sample used.

We normalized any increase in H resulting from publishing longer by dividing H by the number of years since first publication or year publishing as suggested by (Hirsch, 2005). This statistic is known as m. The amount of the variance in m explained by the total number of publications for all years is 17%; by total citations is 40%, by average citations per publication is 28%; by H is 44%; by the number of years publishing is 2%; by the number of papers published from 1996 to 2000 is 23%; by the number of publications from 2001 to 2005 is 21%; and by citations per year is 61%.

Correlation between		**All (N=362)**	**HHMI (N=61)**	**Pioneer (N=25)**
Num Years of Pubs	**H-Index**	0.57	0.62	0.47
Pubs/Year	**H-Index**	0.78	0.64	0.86
Cits/Year	**H-Index**	0.86	0.87	0.88
Total Number of Publications	**H-Index**	0.83	0.88	0.91
Total Number of Citations	**H-Index**	0.91	0.96	0.94
Average Citation per Publication	**H-Index**	0.40	0.65	0.21

Table 3. Correlation of the H-Index with other citation statistics in three different samples

What Bibliographic measures predict future performance?
NIH peer review typically evaluates proposals with project periods lasting 5 years. Thus, NIH peer review attempts to make predictions about the quality of the science produced by an investigator over the duration of the project. This raised the question of what bibliographic measures predict future performance.

To answer this question we acquired bibliographic and citation data for papers published by DBNBR investigators from 1996 through 2000 (96-00) and 2001 through 2005 (01-05) from Thomson Reuters, the publishers of the Web of Science. In this analysis, citation data of papers published from 1996 through 2000 by FY 2006 DBNBR R01 applicants are restricted to citations by papers published from 1996 through 2000, and citation data of papers published from 2001 through 2005 by FY2006 DBNBR R01 applicants are restricted to citations by papers published from 2001 through 2005. Publications included all articles, notes, abstracts, and editorial material archived by Thomson Reuters. Correlations among total number of publications, total number of citations, the H-index, average citations per publication, the average percentile, and c-index were performed. Table 4 shows the correlation coefficients for these data. The correlations shown in Table 4 are for these variables within a 5 year period and how these variables from (96-00) and (01-05) correlate with one another. The sample consisted of 371 investigators out of the original 468 that were provided by Thomson Reuters in which missing data was removed.

	Correlation (N=371)	96-00						01-05					
		Papers	Cites	Avg. Cites/Paper	H-index	C-index	Avg Percentile	Papers	Cites	Avg. Cites/Paper	H-index	C-index	Avg Percentile
96-00	Papers	1.00											
	Cites	0.72	1.00										
	Avg. Cites/Paper	0.10	0.50	1.00									
	H-index	0.80	0.86	0.43	1.00								
	C-index	0.29	0.46	0.65	0.54	1.00							
	Avg Percentile	-0.17	-0.41	-0.82	-0.45	-0.72	1.00						
01-05	Papers	0.81	0.57	0.05	0.62	0.15	-0.07	1.00					
	Cites	0.62	0.74	0.28	0.65	0.23	-0.22	0.71	1.00				
	Avg. Cites/Paper	-0.02	0.13	0.26	0.09	0.12	-0.19	-0.04	0.37	1.00			
	H-index	0.65	0.63	0.21	0.65	0.24	-0.19	0.80	0.85	0.18	1.00		
	C-index	0.07	0.15	0.15	0.14	0.18	-0.10	0.11	0.39	0.73	0.29	1.00	
	Avg Percentile	-0.07	-0.22	-0.30	-0.19	-0.14	0.30	-0.08	-0.37	-0.49	-0.38	-0.56	1.00

Table 4: Correlation between 96-00 and 01-05 using the new data set (N=371)

Within the 96-00 correlation matrix the H-index is strongly correlated with the total number of papers published (r^2 =0.64) and the total number of citations (r^2 =0.74), but is more weakly correlated with the c-index (r^2 =0.29), average percentile (r^2 = 0.20), and average citation per paper (r^2 =0.18). In the case of the 01-05 matrix, the H-index is also strongly correlated with the total number of papers (r^2 =64) and the total number of citations (r^2 =0.72), but more weakly associated with the c-index (r^2 =0.08), average percentile (r^2 =0.14), and average citations per paper (r^2 = 0.03). These statistics also suggest that the associations of the H-index with either total publications or total citations do not vary significantly from one five year period to the next while the association with either the c-index, the citations per publication, or average percentile is much more variable.

Analysis of the correlation matrix for total number of papers shows that the total number of papers are strongly associated with the total citations for both (96-00, r^2 =0.52), and (01-05, r^2 =0.50). A strong association between papers and the H-index as noted in the previous paragraph is seen for both (96-00, r^2 = 0.64), and (01-05, r^2 = 0.64). Much weaker associations are observed in both in the associations of total number of paper with either the average citations per paper, the c-index, and average percentile that range from r^2 =0.00 to r^2 =0.02 for both (96-00) and (01-05) with the exception of c-index with total number of papers (00-96) where r^2 =0.08.

As seen in the previous paragraph the total number of citations is strongly correlated with both the total number of publications and the H-index in both five year periods (96-00) and (01-05), r^2 = 0.74 (96-00) and t r^2 =0.72 (01-05), respectively. The direct relationship between total number of paper and total citations is r^2= 0.52 (96-00) and r^2 =0.50 (01-05). The association of total citations with other bibliographic measures in these two five year periods is much weaker than the relationship between total number of publications or between total number of citations and the H-index. The association of total number of citations with the average citation per paper is r^2 =0.25 and r^2 =0.14, for (96-00) and (01-05), respectively. The association of the c-index with total number of citations per paper is r^2 =0.21 and r^2 =0.15 for (96-00) and (01-05), respectively. The relationship between total number of citations and the average percentile is r^2 =0.17 and r^2 =0.14 for (96-00) and (01-05), respectively. The c-index is inversely proportional to the average percentile because a high percentile is a poor ranking while a high c-index is associated with high-ranking.

The c-index appears to be more strongly associated with the average citation per paper and the average percentile for both five year periods while being more weakly associated with the total number of papers, total citations, or H index. For the periods (96-00) and (01-05) the r^2 for the relationship between the c-index and average citations per paper is 0.42 and 0.53, respectively. A strong relationship was also seen for the c-index with the average percentile (96-00) and (01-05) where r^2 =0.52

and r^2 =0.31. A somewhat weaker relationship is seen between the c-index and H (r^2 =0.29) in the period from 1996 through 2000; r^2 =0.08 in the period from 2001 through 2005). The proportion of the variance in the c-index explained by the total number of papers is 8% in (96-00) and 1% in (01-05). 21 % and 15% of the variance in the c-index is explained by the total number of citations in (96-00) and (01-05), respectively.

Like the c-index the average percentile appears to be more strongly associated with citations per paper and the c-index. 0.52 and 0.32 are the r^2 values for(96-00) and (01-05) for the association between the c-index and the average percentile. . For the periods (96-00) and (01-05) the r^2 for the relationship between the average percentile and average citation per paper is 0.67 and 0.24, respectively. The strength of the association of average percentile with either total number of papers or total number of citations ranged from 0.0 to 0.17.

These data suggest there exists a strong association only among total number of publications, total number of citations, and the H-index, as well as strong relationship only among citations per publication, c-index, and average percentile. Thus, clustering of the 6 variables into two groups may suggest that the two groups are measuring different factors.

As can be seen Table 4., the total number of papers, the total number of citations, and the H-index are the best predictors and most reliable predictors of future performance. 66% of the variance in the total number of papers published from 2001 through 2005 by DBNBR R01 applicants can be predicted by the number of papers they published between 1996 and 2000. 55% of the variance in the total number of citations for paper published from 2001 through 2005 can be predicted by the total number of citations for paper published from 1996 through 2000. 42% of the variance in the H-index in a five year period can predicted the H-index in the next five year period. The bibliographic measures of average citations per publication, c-index, and average percentile in a five year period are not reliable predictors of themselves in the next five year period. The average number of citations per publication in a five year period explains only 7% of the variance of the number of citations per publication in the next five year period. Similarly, the average percentile from 1996 to 2000 explains only 9% of the variance in the next five year period. 3% of the variance in the c-index from 2001 through 2005 is explained by the c-index from 1996 through 2000.

Score and Percentile are not predicted by any bibliographic measure

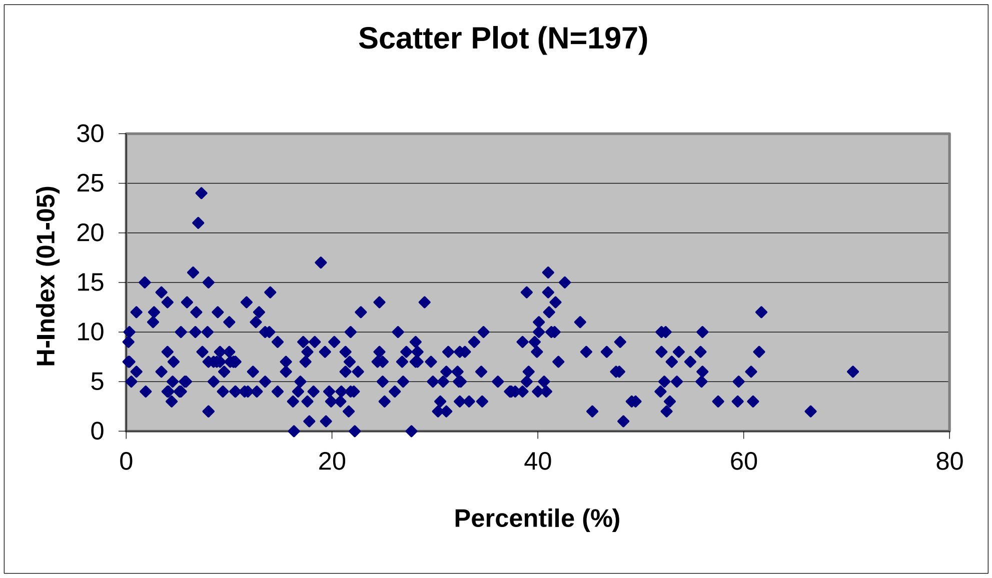
The purpose of peer review of grant applications is to provide advice to NIH program staff about which proposals will produce the best science and to predict which proposals will have the greatest scientific impact. Measures of citation impact are a form of peer review. Highly cited papers are likely to describe methods that will be widely used or a new discovery. The question, then, is whether measures of productivity and citation impact predict the scores and percentiles of grant applications.

197 investigators that were scored are analyzed. The sample size is reduced from 362 to 197 investigators because 165 investigators were not scored. The removal of the 165 investigators did not change the strength of the correlations among the citation impact data compared to the entire sample (data not shown).

While priority score and percentile are highly correlated (r^2=0.81), Table 5 shows that score and percentile by peer review are not predicted by total number of publications, total number of citations, the H-index, average citations per year, average citations per publication, the number of papers published in the previous five years (pubs 01-05), the H-index for papers published and cited only between 2001 to 2005, the total number of citations for papers published and cited only between 2001 to 2005, the average citations per publication for papers published and cited only between 2001 to 2005, the average percentile and cited only between 2001 to 2005, or by the c-index and cited only between 2001 to 2005. Between 0 to 5% of the variance in either score or percentile is explained by these bibliographic measures. The correlations between these bibliographic measures and priority score or percentile except for average percentile is inversely related because a low priority score and percentile are highly ranked while a high bibliographic measure with the exception of average percentile is ranked highly. In Fig 1. the scatter plots of the H-index for all years not just 01 to 05 as a function of priority score or percentile are shown. The scores ranged from 109 to 378 and the percentiles ranged from the 0.2 percentile to the 70.6 percentile. No applications with scores above 200 were funded.

Correlation (N=197)	Percentile	Priority
Percentile	1.00	
Priority	0.90	1.00
Total Pub	-0.17	-0.15
Total Cit	-0.18	-0.15
Avg Cit/Pub	-0.19	-0.14
H-index	-0.23	-0.18
Diff Years	-0.18	-0.10
avg pubs/year	-0.11	-0.12
Num pubs '96-'00	-0.15	-0.13
Num pubs '01-'05	-0.11	-0.11
Delta	0.05	0.02
Cit/Year	-0.17	-0.14
m=H/Years	-0.14	-0.13
Papers 01-05	-0.09	-0.10
Cites 01-05	-0.14	-0.11
Avg. Cites/Paper 01-05	-0.12	-0.08
H-index 01-05	-0.14	-0.12
C-index 01-05	-0.04	-0.03
Avg Percentile 01-05	0.11	0.08

Table 5. Correlation of scored DBNBR 2006 applicants with citation data spanning all years through 2006 and 01-05 data provided by Thomas Reuters N=197.



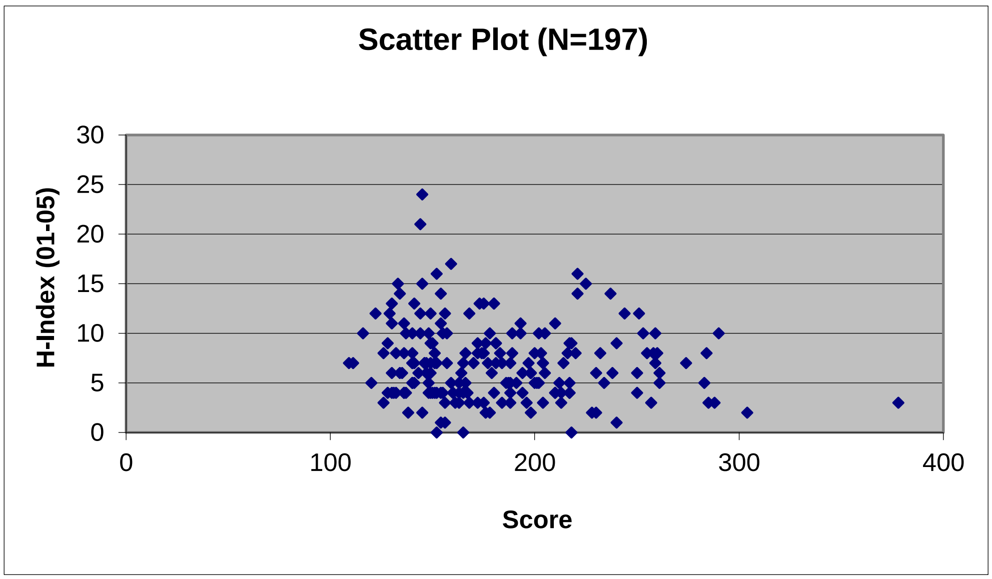


Fig 1. A. Scatter plot of H-index (01-05) as a function of percentile, r^2 =0.05, N=197 (B). Scatter plot of H-index as a function of percentile, r^2 =0.03, N=197. The sample includes all new and competing R01 grant applications submitted to the Division of Basic Neuroscience and Behavioral Research that are listed in the NIH database for awarded, approved by IRG, and withdrawn, regardless of whether the application had been through 1 or two revisions. 165 of the 362 applications were unscored by the IRG.

Funded and Unfunded applications differ in their mean and median H-Index scores and other citation impact measures.

Because unscored applications were not included in the correlation analysis of priority score/percentile with bibliometric measures, creating a possible cut correlation, we determined whether the mean and median H-index scores and citation per publications were different for funded and unfunded applications. Fig 2 shows a histogram plot of the H-index for funded and unfunded investigators for all years through 2006. The histograms for both funded and unfunded show a skewed distribution. Even though there appears to be overlap between the distributions (Fig 2) the mean H-index and standard error of the mean for funded investigators is 28.5± 1.95, N=74 while the mean H-index and standard error of the mean for unfunded is 20.87 ± 0.93 N=288. The median H-index for funded applications is 24 while the median H-index for unfunded applications 17 (Table 6). The differences in the number of years since first publication (years publishing) of funded (28.92 ± 1.42) and unfunded (24.09 ± 0.73) cannot account for the differences observed in H between funded and unfunded DBNBR investigators. H/(years since first publication (m) normalized for year publishing is still different; m is 1.05 ± 0.06 for funded investigators and is 0.89 ± 0.03. A similar result is obtained if you look the H-index restricted to a five year period, 2001 to 2005 (01-05) prior to the review of the applications in which papers published in this period were only cited by paper published during the same five year period. Thus, H here is normalized. The H-index for funded investigator for the period 01-05 is 7.64 ± 0.50 and for unfunded investigators is 5.72 ± 0.21.

Funded investigator did consistently better, on average, than unfunded investigators (Table 6). Funded investigators published more, are more frequently cited, have a greater number of citations per publication or per year and have a higher average percentile ranking of the frequency they are cited compared to investigators. Only in the case of the c-index is no difference found between funded and unfunded investigators. The c statistic measures the difference between the number of citations and the expected number of citations for a journal and then ranks the investigator to produce c.

The probability of being funded as a function of the H-index interval was computed by computing the ratio of funded to unfunded investigators that fell within a given interval of the H-index. As can be seen Fig 2c shows a trend that the probability of funding increases as an investigator's H-index increases. Investigators with an H-index between 0 and 10 have a 6 percent chance of being funded. Investigators falling with H-indexes between 11 and 30 have probabilities that are close 19%, the funding rate for DBNBR for FY 2006. Investigators with H-index between 31 and 50 had a 30 to 40 percent chance of being funded. Investigators with H-scores in the range of 51 to 60 had a success rate around 22%. An H-index between 61 and 70 had a 55% chance of being funded. Investigators with H-index scores between 71 and 80 had about a 100% chance of being funded. The success rate of funding declined above

	Mean ± S.E.M Funded N=74	Mean ± S.E.M Unfunded N=288	Median Funded N=74	Median Unfunded N=288
Total number of publications	132.59 + 14.12	94.45 + 7.77	99	54
Total number of citations	4012.54 + 586.33	2463.76 + 302.70	2448.50	1121.5
Average citations per publication	29.92 + 2.52	23.86 + 1.01	22.27	20.08
H-index	28.65 + 1.95	20.87 + 0.93	24.00	17
Years Publishing	28.92 + 1.42	24.09 + 0.73	28.50	22
Average number of publications per year	4.28 + 0.31	3.43 + 0.18	3.51	2.44
Citations per Year	127.70 + 16.13	86.65 + 8.11	88.04	50.16
m= H/years publishing	1.05 + 0.06	0.89 + 0.03	0.91	0.83
Total number of publications 01-05	30.05 + 2.86	22.60 + 1.25	24.00	16
Total number of Citations 01-05	267.30 + 41.74	160.23 + 12.21	158.50	97
Average citation per publication 01-05	9.10 + 0.88	7.67 + 0.58	6.15	5.51
H-index 01-05	7.64 + 0.50	5.72 + 0.21	7.00	5
C-index 01-05	0.51 + 0.02	0.51 + 0.02	0.49	0.48
Avg Percentile 01-05	59.42 + 1.59	63.97 + 0.82	59.81	65.23

Table. The means and standard errors of the means for citation impact scores for funded and unfunded DBNBR FY2006 applicants for all years publishing and for citations restricted to papers published from 2001 to 2005.

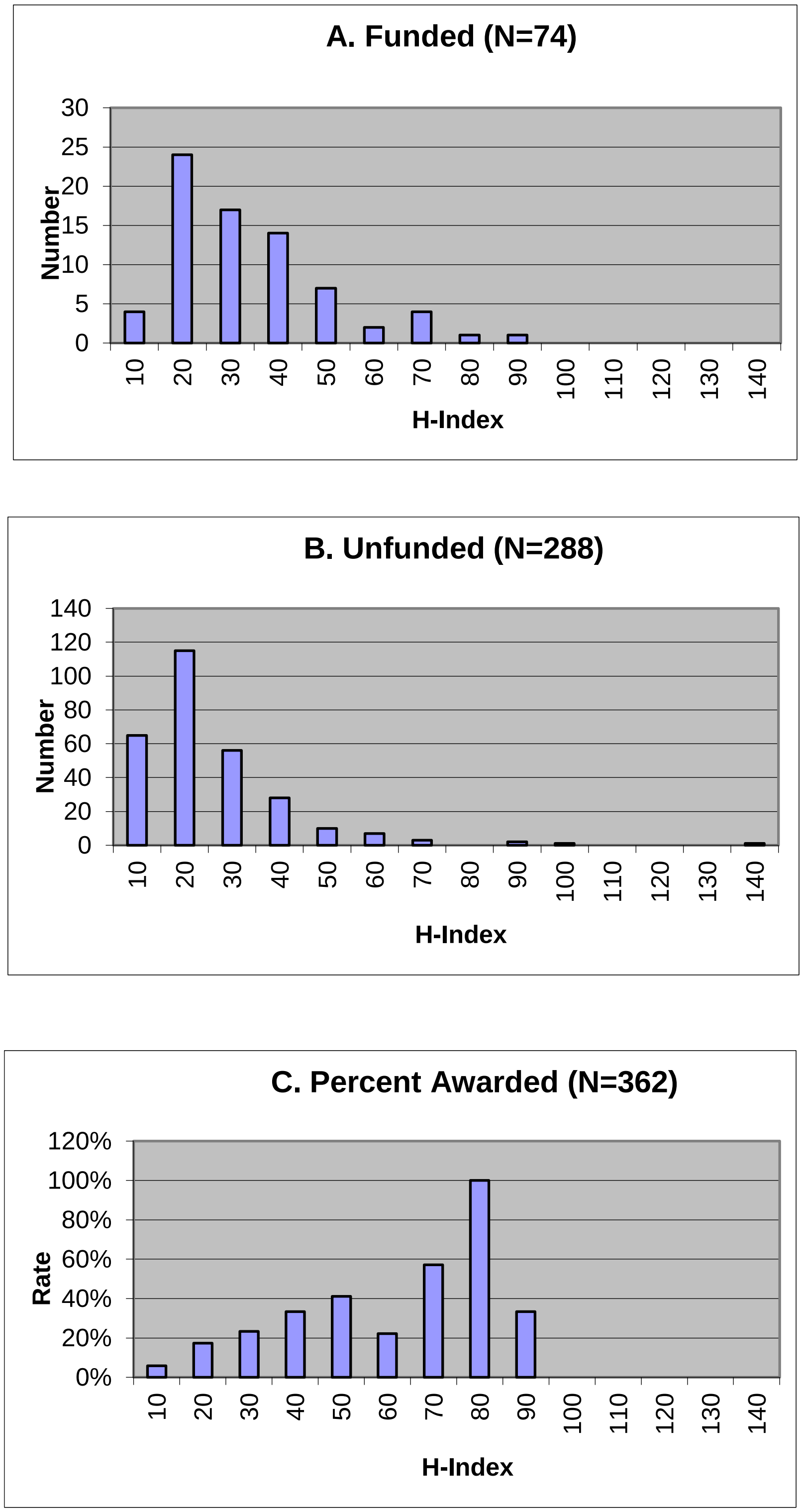


Fig 2. A) Histogram distribution of H-index for funded investigators. B) Histogram distribution of H-index for unfunded investigators. C) Percent awarded by H-index.

Comparison of the means for H-index, citation per publication, citations per year, and publications per year for funded, unfunded, HHMI investigators, and Pioneer awardees

The next question we asked is whether citation data has any validity as a measure of quality of an investigator. Because the 61 HHMI Neuroscience investigators include Nobel Laureates and members of the National Academy of Science we tested the hypothesis that HHMI investigators would have greater impact as measured by the H-index, citations per year, publications per year. The means for HHMI neuroscience investigators, Pioneer awardees, Funded DBNBR R01 applicants, all DNBNR R01 applicants, and unfunded DBNBR applicants are displayed in Fig 3. The means and s.e.ms for the H-index is 48 $\pm$ 3.4 for HHMI, 28.65 $\pm$ 1.95for Funded DBNBR Awardees, 25 $\pm$ 2.8 for (Pioneer awardees), 22.46 $\pm$ 0.86 for all DBNBR R01 applicants both Funded and unfunded), and 20.87 $\pm$ 0.93 for unfunded DBNBR R01 applicants. Thus, the rank order for the H-index from greatest to least is HHMI > DBNBR funded R01 applicants> Pioneer Awardees > all DBNBR R01 applicants> unfunded DBNBR applicants.

The H-index increases with years of publishing. To normalize the H-index, which increases with years publishing, the H-index is normalized by dividing the H index by the years since investigator first published to create the statistic m. As seen in figure HHMI investigator perform better (mean= 1.82 +0.10) than Pioneer (mean= 1.13 $\pm$ 0.10), funded (mean= 1.05 $\pm$ 0.06), and unfunded investigators (mean =0.89 +. 0.03). Only does the standing of Pioneer investigators relative to funded and unfunded DBNBR investigator increase as compared to using the H Index.

A similar rank order is seen for average citations per publication: HHMI (70 $\pm$ 4.5), Pioneer Awardees (35 $\pm$ 2.8), Funded DBNBR Awardees (29.92 $\pm$ 2.52),all DBNBR Awardees (25.10 $\pm$ 0.9); Unfunded (23.86 $\pm$ 1.01) as well as for citations per year 398 $\pm$ 41.2 (HHMI), 138 $\pm$ 27.27 (Pioneer Awardees), 127.70 $\pm$ 16.13 (Funded DBNBR R01 Applicants), 95.04 $\pm$ 7.29 (All Funded) and 86.65 $\pm$ 8.11 (Unfunded DBNBR Applicants).

HHMI investigators published on average one more paper per year (5.44 $\pm$ 0.42) than either the Pioneer Awardees (mean = 4.05 $\pm$ 0.70 publications per year), Funded DBNBR R01 (mean = 4.28 $\pm$ 0.31 publications per year), applicants, all DNBNR R01 applicants (mean = 3.6 $\pm$ 0.16 publications per year),, or unfunded DBNBR applicants (mean = 3.43 $\pm$ 0.18 publications per year). Thus, there is not a great difference in the publication rate among the groups

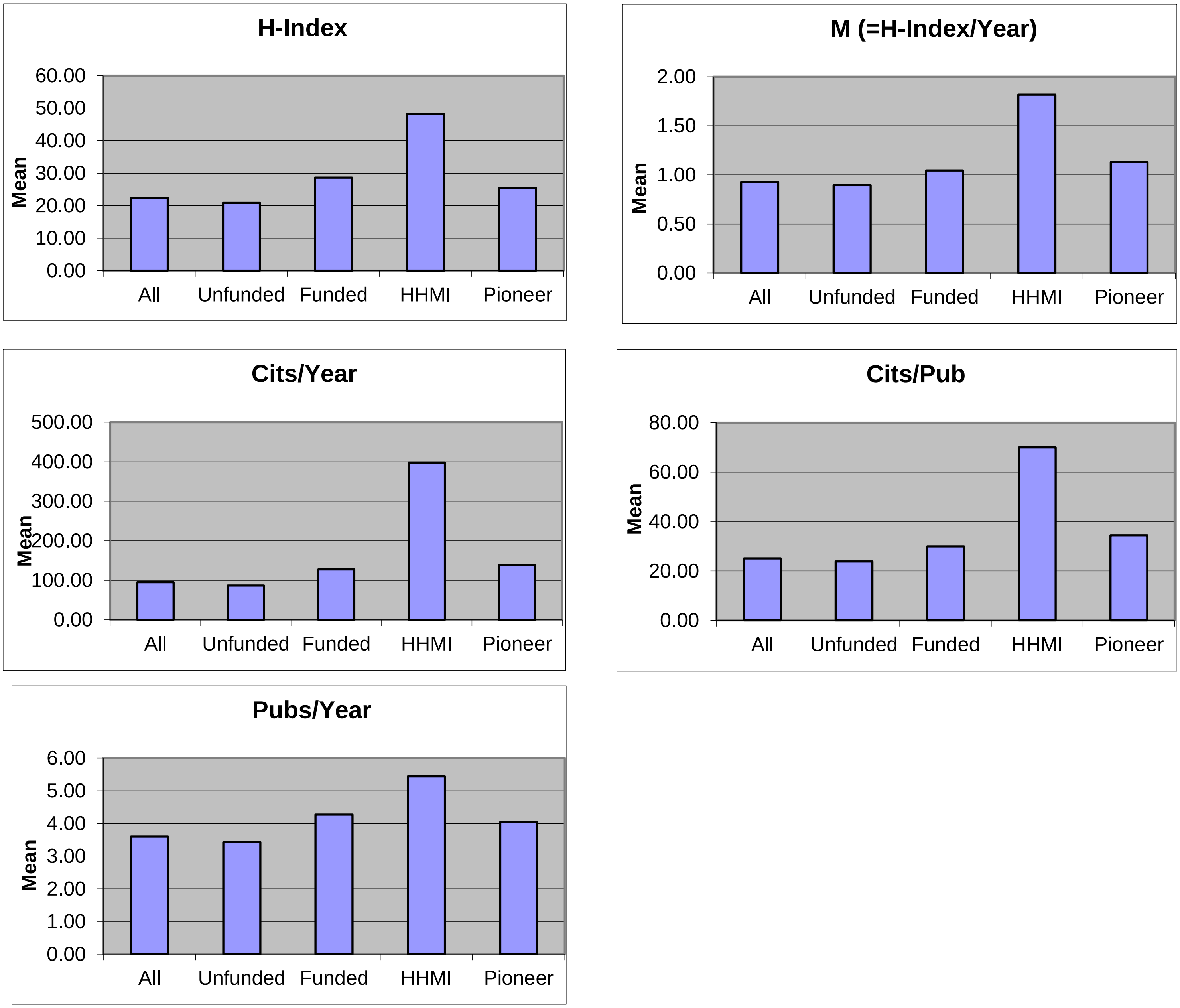
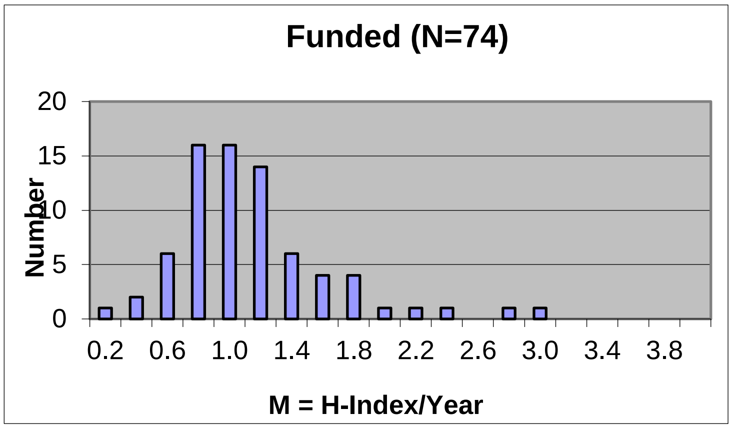


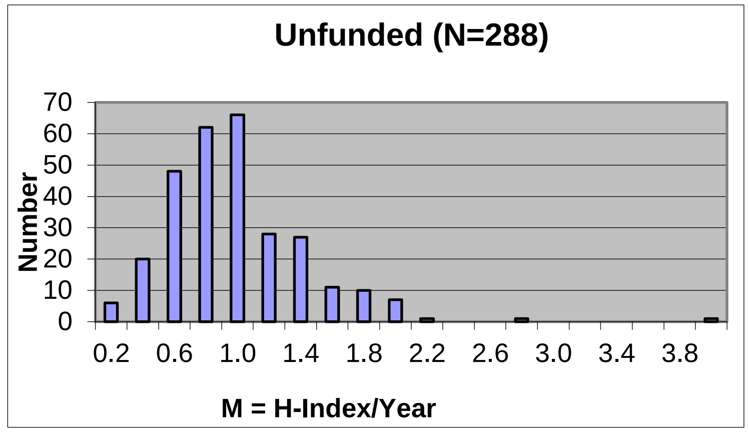
Fig 3. Histograms comparing means for H-index, citations per publication, citations per year, and publications per year for All DBNBR R01 Applicants (N=362), Funded DBNBR R01 Applicants (N=74, Unfunded DBNBR Applicants (N=288, 61 HHMI neuroscientists, and 25 Pioneer Awardees.

Correlation of citation impact score with peer review criterion score for investigator and weighting of criterion scores in predicting average preliminary score

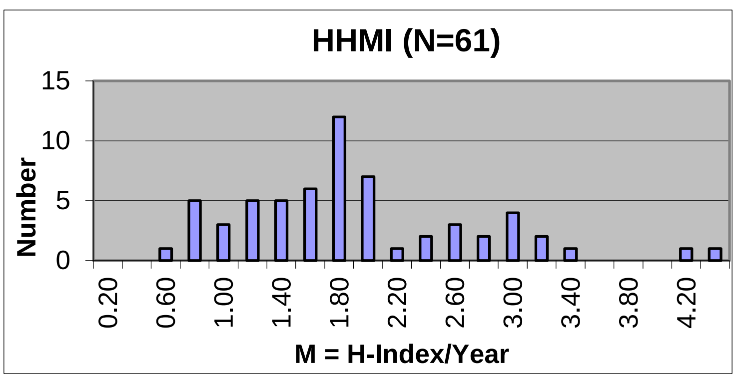
The lack of correlation of impact factors and peer review may be due to a cut correlation or other factors weighted in the review such as significance, approach, innovation, and environment. The recent introduction of reporting the criterion score for significance, investigator, approach, innovation, and environment in review now permits an analysis of whether citation impact score is correlated with the evaluation of investigator and how each criterion is weighted in review. In addition, the inter-rater reliability can be estimated.

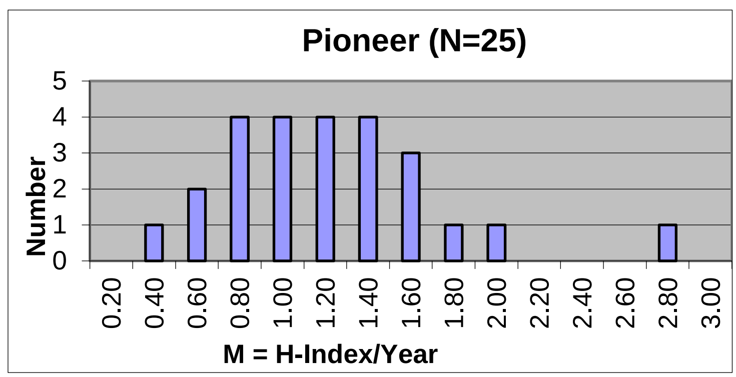
Fig4. Histogram of M= H-index/years since first publication for Funded FY 2006 DBNBR





To conduct this analysis 71 R01 application reviewed by Molecular Neuropharmacology and Signaling Study Section were examined. Of the 71 applications reviewed, 38 were reviewed in June 2009 and 33 were reviewed in October 2009. 46 applications were assigned a priority score and corresponding percentile, while 25 were not discussed. 20 out of the 71 applications reviewed were first resubmissions (A1) and 6 applications were second resubmissions (A2). Finally, 53 of the 71 applications were new applications and 18 were competing renewals.





As a measure of internal agreement within peer review, correlations between reviewer's criteria scores were analyzed within each separate review category. Table 1 illustrates correlations of the 5 criteria scores between the individual reviewers of each application. Mean correlations for each criteria score and all reviewers were subsequently calculated for each review criteria.

	reviewer {1,2} (N=73)	reviewer {1,3} (N=71)	reviewer {2,3} (N=71)	mean
significance	0.13	0.22	0.37	0.24
investigator	0.24	0.37	0.49	0.37
innovation	0.49	0.39	0.56	0.48
approach	0.39	0.37	0.42	0.39
environment	0.37	0.32	0.49	0.39

Table1: Correlation between reviewer's criteria scores for MNPS applications Jun/Oct 2009 (n=71 and n=73).

Between reviewers, scores for significance (r^2= 0.06), investigator (r^2= 0.14), innovation (r^2= 0.23), approach (r^2= 0.15), and environment (r^2= 0.15) were weakly correlated with each other. Reviewers were most in agreement with regards to their evaluation of innovation within an application and least in agreement on the significance of an application; however, it appears agreement between reviewers was weak overall.

	Reviewer 1 Prelminary	Reviewer 2 Prelminary	Reviewer 3 Prelminary
Reviewer 1	1.00		
Reviewer 2	0.57	1.00	
Reviewer 3	0.54	0.52	1.00

Table 2: Correlation between reviewers' preliminary scores (n = 71).

Correlations between reviewers' preliminary overall scores for all applications are shown above in Table 2. Preliminary overall scores were moderately correlated between all three reviewers (r^2= .32, .29, .27, respectively), indicating agreement in the overall consideration of an application.

A two-way analysis of variance was performed on average preliminary score and each reviewer's initial score to examine the measure of error associated with a reviewer's initial overall score. The standard deviation across reviewer's initial scores for each application was 1.17.

n=217	environment	significance	investigator	innovation	approach
environment	1.00				
significance	0.49	1.00			
investigator	0.62	0.55	1.00		
innovation	0.59	0.63	0.61	1.00	
approach	0.50	0.60	0.57	0.61	1.00

Table 3: Correlations between criteria scores.

Table 3 demonstrates the inter-correlated nature of each of the five criteria scores. All scores are moderately associated with each other, with the highest correlations between innovation and investigator (r^2=0.38), approach (r^2=0.37), and significance (r^2=0.39) and between environment and investigator (r^2=0.38).

	investigator rev1 (n=72)	investigator rev2 (n=72)	investigator rev3 (n=70)	Average (n=68)	Final Score (n= 47)
h-index	-0.34	-0.34	-0.33	-0.37	-0.39
m	-0.33	-0.35	-0.31	-0.40	-0.13
cit/pub	-0.18	-0.21	-0.17	-0.33	-0.30
h-index 04-08	-0.34	-0.47	-0.48	-0.42	-0.23

Table 4: Correlations between bibliographic measures of impact and average, final overall and investigator scores.

The h-index is measure of publication impact that is defined as the number of papers published by an investigator that are cited at least h-times, i.e. an h-index value of 10 indicates 10 papers cited at least 10 times. M is defined as the h-index divided by the number of years an investigator has published. Cit/pub is the average number of citations per publication by each investigator. Finally, h-index 04-08 is the h-index limited to the years 2004-2008. All four measures were inversely correlated with the investigator criterion score, average preliminary score, and final overall score. H-index was mildly correlated with reviewers' investigator criteria score (r^2= -0.12, -0.12, -0.11), average preliminary (r^2=-0.14), and final overall score (r^2= -0.15). M was also modestly correlated with investigator (r^2=-0.11, r^2= -0.12, and r^2=-0.10), average preliminary (r^2=-0.16) and with final overall scores (r^2=-0.13). Citations per publication was very weakly correlated with reviewers' investigator criterion scores (r^2=-0.03, -0.04, -0.03) and only very slightly more strongly associated with average preliminary and final overall scores (r^2= -0.11 and r^2=-0.09, respectively). H-index over the period of 2004-2008 was modestly associated with the investigator criterion score (r^2 = 0.12, 0.22, 0.23), Average preliminary score (r^2 =0.18), and final score (r^2= 0.05). The consistency in correlations between h-index and m and investigator criterion score across reviewers indicates reviewers similarly weight measures of publication impact in their consideration of investigator criterion score.

	Significance rev1	Investigator rev 1	Innovation rev 1	Approach rev 1	Environment rev1
Average preliminary score (n=69)	0.45	0.43	0.51	0.52	0.40
	significance rev 2	investigator rev 2	innovation rev 2	approach rev 2	environment rev 2
Average preliminary score (n=69)	0.48	0.54	0.62	0.69	0.52
	significance rev 3	investigator rev 3	innovation rev 3	approach rev 3	environment rev 3
Average preliminary score (n=67)	0.61	0.59	0.60	0.72	0.61
Mean	0.51	0.52	0.58	0.64	0.51

Table 5: Correlations of criteria scores with average preliminary scores.

Table 5 displays correlations between each reviewer's individual criteria scores and the average preliminary score for each application. Across reviewers, criteria scores were moderately correlated with average preliminary score. As shown by mean correlation between criteria score and average preliminary score, approach was at most moderately correlated with average score (r^2= .41), followed by innovation (r^2=0.33), investigator (r^2=0.27), significance (r^2=.26), environment (r^2=0.26).

Partial correlation analysis was performed between the five criteria scores and average preliminary score to determine the association between each criteria score on average preliminary score while holding all other variables constant (Table 5).

Average Preliminary Score (n=205)	significance	investigator	innovation	approach	environment	sum
r	0.07	0.05	0.16	0.37	0.13	
r^2	0.005	0.002	0.026	0.135	0.018	0.186

Table 6: Partial correlation analysis of criteria score and preliminary average score (n=205).

When controlling for all other criteria variables, preliminary average score was only negligibly correlated with scores for significance (r^2= 0.005), investigator (r^2= 0.003), innovation (r^2= 0.03), and environment (r^2=0.02). Approach was slightly stronger in its association with average preliminary score (r^2= 0.14). Individually, criteria scores for significance, investigator, innovation, approach and environment only individually explain 0.5%,0 .2%, 2.6%, 13.5%, and 1.8% of the variance in average preliminary score, respectively.

Multiple regression was performed between average preliminary score and the five criteria scores pooled among all three reviewers: significance, investigator, innovation, approach, and environment. The order of in which the variables were input into the model was determined by the squared partial correlation shown in Table 6; variables were entered according to the amount of variance explained, from greatest to least. The overall model was significant ($p < 0.0001$). The equation for the full model is as follows:

Average Preliminary Score = .296*Approach + 0.166* Innovation + 0.060*Investigator + 0.060*Significance +0.146*Environment + 1.70

Within the model, only innovation ($p = 0.021$) and approach ($p < 0.0001$) were significantly associated with average preliminary score. Environment was borderline significant ($p = 0.0577$). As the partial correlation analysis above demonstrated, both innovation and approach had the strongest association with average preliminary score, which is in agreement with results of the regression model. Overall, the model could explain 48% of the variance in average preliminary score (Adjusted R^2 = 0.475, Unadjusted R^2= 0.488).

Squared partial correlations were calculated for each independent variable in the model, as illustrated by Table 7.

n=205	significance	investigator	innovation	approach	environment	Sum
r^2	0.005	0.002	0.026	0.135	0.018	0.186

Table 7. Type II SS Squared Partial Correlations calculated within the regression model for each independent variable.

Squared partial correlations were calculated for each variable within the model using Type II sum of squares, as opposed to Type I sum of squares. Type II sum of squares is computed by assuming all the independent variables are present in the model and controlling for the influence of all other effects of an equal or lower degree, whereas Type I sum of squares is calculated sequentially controlling for each variable as they are added to the model. As indicated in the table, approach and innovation explain the largest portion of the variance (14% and 3%, respectively), followed closely by environment (2%), which matches the results of the independent partial correlation analysis illustrated above in Table 6. In addition, these results are consistent with the results of the multiple regression, with approach and innovation being the only significant ($p<0.05$) predictors of average score.

Squared semipartial correlations were calculated for each independent variable in the model, as shown in Table 8.

n=205	approach	innovation	environment	significance	investigator	Sum
r^2	0.41462	0.0543	0.01494	0.00286	0.00119	0.488

Table 8. Squared semipartial correlations calculated within the regression model for each independent variable.

Squared semi partial correlations were calculated using Type I sum of squares to examine the proportion of variance explained by each variable. As explained above, this takes into account sequential addition of variables into the model. From examining the table, it is evident that approach accounts for the vast majority of explained variance in the multiple regression model (r^2=0.41), with innovation (r^2=0.05) and environment (r^2=0.015) accounting for nearly all the remainder. Both significance (r^2=.002) and investigator (r^2=0.001) explain a negligible amount of the variance in the model. The sum of the squared semipartial correlations is equal to the unadjusted R^2 of the model (R^2=0.488). As demonstrated by the model equation, the approach criterion score is the largest predictor of average preliminary score, whereas the investigator criterion score is weighted the least. These results are in agreement with both the partial correlation analysis (Table 6) and semipartial correlation analysis (Table 8) demonstrated above.

Discussion

Our results suggest that the H-index and citations per publication are valid measures of quality. HHMI neuroscience investigators, many of them, Nobel Laureates, and members of the National Academy Sciences outperformed DBNBR investors on these performance criteria. It could be argued simply that the larger H-index is the result of seniority and being well funded. However, when the H-index is normalized to the number of years since first publication significant differences are still seen in m between DBNBR R01 FY 2006 applicants and HHMI.

The results presented here show that the number of papers, the total number of citations, and h-index in a five year period are good predictors of performance for those variables for the next five years while average percentile, citation per publication are weaker predictors of performance. The correlation coefficient of number for number of publications published by an applicant during 1996 to 2000 with the number of publications published by an applicant from 2001 to 2006 0.8 is similar to the numbers reported by (Abrams, 1991) for 11 members of the 1988 NSF Ecology panel and 45 ecology scientists referenced from a 1973 ecology text book. Abrams reports that the number of publications during 1980 to 1984 was strongly associated with the number of publications published from1975 to 1979 the 11 members 1988 NSF Ecology panel (r=0.775). In the case of the 45 scientists referenced in the ecology textbook, the correlation between papers published from 1970-1974 with the number published from 1975-1979 is r=0.794 after giving lower weights to multiple authored papers, an analysis not conducted in the present paper. (Hirsch, 2007) reports that the correlation coefficient for publications in one sample of 50 physicists and another sample of 29 physicists in a 12 year period for predicting the total number of publications the subsequent 12 years is r=0.43 and r=0.50, respectively, a more modest association than reported here or by (Hirsch, 2007)

The strong association between total number of citations in a five year period predicting the total number of citations for the next five years (r=0.74) is similar to that observed (Abrams, 1991). Abrams reports a high correlation (r=0.875) for the total number of citations of papers published during a five year period from 1973 to 1977 with the total number of citations of papers published during a five year period from 1980-1984 by the 11 members of the 1988 NSF Ecology Panel. A weaker association is reported is by Hirsch(Hirsch, 2007) of r=0.53 and r=0.43 for total number of citations during a 12 year period predicting the next 12 year period for the two groups of physicists sampled.

The H-index is also a good predictor of itself. The correlation of h for 1996 to 2000 with h for 2001 to 2005 is r=-0.65. This is in good agreement with (Hirsch, 2007) who reports a correlation of r=0.61 for his sample of 50 physicists selected from the journal *Physical Review B: Condensed Matter and Material Physics*. In contrast, the results reported here for the average number of citations per publication in a five year period of time appears to be a poor predictor of citation impact of the citations per publication for the next five years (r= 0.26). This is similar to that reported by Hirsch (r=0.23). The weak correlation of average number of citations per paper for two consecutive five year periods may be explained by an investigator publishing more papers with the same total citations in the next five year period or publishing fewer papers with the same or greater number of citations. The combined variance of the two variables of total citations and total number of publications may make this statistic a poor predictor of performance. This may also be true of the c-index and the average percentile.

The limitation of the H-index is that two investigators can have the same H-index but one investigator could have a larger total number of citations for the same H number of papers. In other words, the H-index does not account for the number of papers in excess of h^2 (Bornmann et al., 2008a;Hirsch, 2007;Zhang, 2009). The strong association of total publications with total citation and h-index but not as much with citations per publication, c-index, and average percentile shown here may suggest that these bibliographic measures may be measuring two factors. The H-index, total citations, and total publications may be a measure of productivity while average number of citations per publication, the c-index, and average percentile may be measures of impact. (Bornmann et al., 2008a;Hirsch, 2007) identified Quantity of Productive Core and Impact of productive core as a two factor solution to a factor analysis of 9 bibliographic indices. The H-index and m lie along the dimension of the productive core while indices such as a-index (the average number of citations of the h number of papers or the ar-indices (the square root of the sum of the total number of citations for h number of papers divided by years since first publishing) appear to measure the impact of the productive core. It will be of interest in the future to determine whether the a-index, the ar-index, and e^2 index (total number of citations for h number of papers – h^2, a measure of excess citations) (Bornmann et al., 2008a;Hirsch, 2007;Jin et al., 2007;Zhang, 2009) are better predictors of impact than citations per publication, the c-index, or average percentile for subsequent five year period.

Our data shows a weak to modest association of the H-index, m, and average citations with average score, final score, and the criterion score for investigator. The amount of the variance explained by any bibliographic measure for the five year period preceding the review of FY 2006 DBNBR applicants with score was not greater than 5%. The amount of variance in average preliminary score explained by average citation per publication or h was not greater than 16%. The lack of a strong correlation between bibliographic indices and scores of grant applications or investigator has been reported previously. Cole et al (1981a, b) reported that a between 1 to 16% of the variance in NSF panels for biochemisty, meteorology, or ecology by either number of citations or publications. The correlation between publication track record of 2001 applicants as assessed by peer review groups for the Australian National Health and Medical Research council and number of publications or total citations was 0.375 and 0.327 (Nicol et al., 2007). The lack of a strong correlation between bibliographic measures and their own evaluation may be explained by the fact that reviewers do not have sufficient data to evaluate the track records of investigators because reviewers are only given a list of publication and not bibliographic data. Analysis of the criterion scores for investigator show the scores to be bunched up and not spread out.

Although a weak to moderate association between peer review and bibliographic measure are observed, funded DBNBR FY 2006 investigators had higher citation impact scores than DBNBR FY 2006 unfunded investigators. The probability of receiving an award was associated with the H-index. These observations are consistent with other studies. Bornmann and colleagues (Bornmann et al., 2008b) (Bornmann et al., 2008c) found that funded investigators reviewed by the European Molecular Biology Organization (EMBO) for the EMBO Long-Term Fellow and Young Investigator programs had larger number of citations, greater number of publications, and a larger H-index than unfunded applicants. (Bornmann and Daniel, 2007) also reports that the H-index of successful post-doctoral fellows applicant funded by Boehringer Ingelheim Fonds based on peer review was higher than unsuccessful applications. Van Raan (van Raan, 2006) reports that h index scores tend to be found among the 147 research groups in chemistry in the Netherlands that are judged by peers to be excellent or good than in groups judged to be satisfactory.

If the goal of NIH in enhancing peers review is to place more weight on the track record of the investigator and increase and make the review more transparent, then NIH has failed. Our multiple regression and partial correlation analysis of MNPS study section score for June and October 2009 review suggest that reviewers give the greatest weight to the criterion scores of approach and innovation and give the least weight to significance and to the investigator in arriving at their preliminary scores. 96% of the 48% variance in average score that is explained by the criterion scores is accounted by the approach and innovation in our regression model. Allowing investigators to weigh the different criterion score in any manner to arrive at their preliminary score is not transparent. Investigators, whose applications are not discussed, are not provided preliminary scores. These applicants cannot judge what preliminary score would have been received because the weighting is arbitrary and not all of the variance is accounted by the criterion scores. The lack of the preliminary scores in the summary statement makes evaluation of the review more difficult by NIH program staff and access to such data is at the discretion of the Scientific Review Officer (SRO). Moreover, because reviewers do not weight heavily the track record of the investigator or accurately judge the track record based on bibliographic measures many of the most productive investigators that applied in FY 2006 went unfunded.

The fact that peer review is able to discriminate on average more productive investigators even though bibliographic measure are weakly correlated and that environment, approach, and significance are weighed more heavily than investigator, significance, and innovation suggest that these variables are indirectly correlated with citation impact. Investigator who write well written applications in which the approach is methodologically sound are likely to be the ones who are able to write manuscripts well and think clearly. Moreover, stronger investigators who will produce science with greater impact are likely to be at scientific institutions with other investigators that have greater scientific impact as measured by bibliographic measures(Hendrix, 2009;van Raan, 2006).

The lack of inter-reviewer reliability is of great concern. Our analysis shows that inter-reviewer reliability for the criterion scores are weak with the best correlation coefficient of r=0.39. The inter-reviewer reliability for the preliminary score is a little better at around r=0.55 and similar to that reported by Cole (Cole et al., 1981) for NSF peer review study section in chemical dynamics, economics and solid state physics. The standard deviation for the score of assessed by the three reviewers is 1.17. Thus, computing the 90 % confidence interval for a score of 3 is 1.03 to 4.97. The idea that the variance is reduced by the three reviewers discussing the application and then having the study section vote their conscience is flawed. The assertiveness and articulateness of a reviewer may pull the final preliminary scores of the other reviewers. Because many members of the study section have not read the application and may not have expertise, their vote is not an independent assessment of the application and is entirely influenced by the arguments of the three reviewers. Thus group dynamics artificially reduces the variance and may produce a situation of the story of the king's clothes. A larger number of independent reviewers would increase the accuracy of the review but may not be practical because of the large number of proposal that need to be reviewed and the effort to find more reviewers to independently review the proposals.

Even if the accuracy of peer review is flawed as currently constituted, the peer review system has great merit. The application process assists scientists in organizing their thought and developing a detailed plan for future experiments. Peer review provides invaluable feedback that helps investigators improve their science, something that funding decision solely based on bibliographic metrics would not provide. Furthermore, citation impact measures may miss important discoveries because the field is a nascent field that is small with few scientists citing the work(2009;Cole et al., 1981) requiring experts to recognize the significance of the work. In addition, different bibliographic indices may measure different factors (2009;Berghmans et al., 2003;Cole et al., 1981). The inter-reviewer reliability might be increased by providing training sets of applications to reviewers. It will be of interest to determine whether reviewers with high citation impact score make better reviewers as judge by their ability to accurately assess citation impact of an applicant.

The review process will be greatly strengthened by combining citation impact measures with peer review. By combining citation impact data with the current peer review system the accuracy of peer review can be improved without overburdening the system to obtain more independent reviews and at the same time take advantage of the strengths of peer review just mentioned. To achieve this objective, the ranking of H-index scores would be ranked from 1 to 9 and weighted by program with the peer review score to produce a weighted combined score. . The effect of placing different amount of weights on bibliographic measures on award outcome needs to be evaluated. Given that more than 50 percent of the variance is not explained by all the criterion variables it is not unreasonable to give the h-index a weight of 40% of the score when combined with the peer review score. The predictive value of different citation impact measures, other than the ones used here need to be evaluated as well. In undertaking this effort care always needs to be taken to ensure that citation impact scores are compared within fields and not across fields.

Reference List

Anonymous (2009) Experts still needed. Nature 457:7-8.

Abrams PA (1991) The Predictive Ability of Peer-Review of Grant Proposals - the Case of Ecology and the United-States-National-Science-Foundation. Social Studies of Science 21:111-132.

Berghmans T, Meert AP, Mascaux C, Paesmans M, Lafitte JJ, Sculier JP (2003) Citation indexes do not reflect methodological quality in lung cancer randomised trials. Annals of Oncology 14:715-721.

Bornmann L, Daniel HD (2007) Convergent validation of peer review decisions using the h index - Extent of and reasons for type I and type II errors. Journal of Informetrics 1:204-213.

Bornmann L, Mutz R, Daniel HD (2008a) Are there better indices for evaluation purposes than the h index? a comparison of nine different variants of the h index using data from biomedicine. Journal of the American Society for Information Science and Technology 59:830-837.

Bornmann L, Wallon G, Ledin A (2008b) Does the committee peer review select the best applicants for funding? An investigation of the selection process for two European molecular biology organization programmes. PLoS One 3:e3480.

Bornmann L, Wallon G, Ledin A (2008c) Is the h index related to (standard) bibliometric measures and to the assessments by peers? An investigation of the h index by using molecular life sciences data. Research Evaluation 17:149-156.

Cole S, Cole JR, Simon GA (1981) Chance and consensus in peer review. Science 214:881-886.

Garfield E (2006) The history and meaning of the journal impact factor. Jama-Journal of the American Medical Association 295:90-93.

Hendrix D (2009) An analysis of bibliometric indicators, National Institutes of Health funding, and faculty size at Association of American Medical Colleges medical schools, 1997-2007. (vol 96, pg 324, 2008). Journal of the Medical Library Association 97.

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci U S A 102:16569-16572.

Hirsch JE (2007) Does the H index have predictive power? Proc Natl Acad Sci U S A 104:19193-19198.

Jeang KT (2007) Impact factor, H index, peer comparisons, and Retrovirology: is it time to individualize citation metrics? Retrovirology 4.

Jeang KT (2009) The importance of individualized article-specific metrics for evaluating research productivity. Retrovirology 6.

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

Nicol MB, Henadeera K, Butler L (2007) NHMRC grant applications: a comparison of "track record" scores allocated by grant assessors with bibliometric analysis of publications. Med J Aust 187:348-352.

Pendlebury DA (2009) The use and misuse of journal metrics and other citation indicators. Archivum Immunologiae et Therapiae Experimentalis 57:1-11.

Seglen PO (1997) Why the impact factor of journals should not be used for evaluating research. British Medical Journal 314:498-502.

van Raan AFJ (2006) Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. Scientometrics 67:491-502.

Zhang CT (2009) The e-index, complementing the h-index for excess citations. PLoS One 4:e5429.