大学評価　第３号　平成15年９月（論文）
[大学評価・学位授与機構　研究紀要]

ビブリオメトリクスによるピアレビューの支援可能性の検討
理学系研究評価の事例分析から

Is Bibliometrics Useful to Support the Peer-review?
A Case Study of NIAD's Research Evaluation in Science

林　隆之
HAYASHI Takayuki

Research in University Evaluation, No.3（September, 2003）[the article]
The Journal of University Evaluation of National Institution for Academic Degrees and University Evaluation

ビブリオメトリクスによるピアレビューの支援可能性の検討
理学系研究評価の事例分析から

林　隆之[1]

１．はじめに

研究評価において，研究活動の成果をその学問的な質から評価する場合には，評価対象と同じ研究分野の専門家に判断を委ねる「ピアレビュー」が主たる方法であることは各国で共通する[2]。これは，研究活動は高度に専門的な知識を要する活動であるため，その成果の価値を適切に判断できるのは同じ専門知識を有している研究者のみであるという前提に基づくものである。ピアレビューは科学者共同体が自己規制によりその質を維持するメカニズムであるとされ，研究活動という社会的営みを支える基盤的システムとして17世紀以降に制度化されてきた（Merton 1973, Chubin and Hackett 1990)。そもそもピアレビューは学術雑誌への投稿論文の掲載可否を判断するための方法として生じたが, 現代では研究活動への資金の配分, さらには個別の活動を越えた大学や研究機関の評価にもピアレビュー方式がとられることは多い。

しかし，ピアレビューにも様々な問題があることは指摘されている（例えば Chubin and Hackett 1990, Kostoff 1994)。その最大のものは，ピアレビューは原則的に評価者（レビューアー）の判断であるために，評価者の主観による意識的・無意識的なバイアスが入りうることである。具体的には，既存の分野を守ろうとする保守的な傾向（オールド・ボーイ・ネットワーク), 若い研究者や新参者の過小評価, 著名な研究者や機関に高評価を与えてしまう「ハロー効果」，個人的・組織的なえこひいきなどがある。さらには，そもそも「優れた研究とは何であるか」に対する考え方の違いは評価結果に大きく影響する。また別の問題として，ピアレビューは分野ごとの専門家による評価であるため，分野をまたがる学際的な研究の判断は行いづらく，異なる分野間での比較を行うことも困難である。そのため，例えば分野を超えて資金配分を行う場合にはピアレビューはうまく機能しにくいことになる。また，評価者一人が実際に評価を行える数は限られるため，少ないサンプル数の中での比較に基づく判断となり，評価対象群の構成による影響を受けやすくなる。さらに，評価対象の数が多く，分野が多様である場合には，全ての分野を網羅する評価者を揃えることも難しく，評価者が評価にかける時間や資金も莫大なものになる。

このようなピアレビューの問題に対して，その質を向上するためには様々な方法がとられる[3]。例えば評価基準の詳細な明示化と評価者の間でのコンセンサス形成，評価者の分野や年齢のバランスを考慮した選出，評点の集計方法の改善などが挙げられる。そのような改善策

1　大学評価・学位授与機構　評価研究部　助手

2　研究活動の成果を，学問的な質のみでなく，その経済・社会的効果から評価する場合には，ピアだけでは評価を行えない場合も多い。その場合には，当該研究分野の専門家以外をも含めた専門家パネル方式や，経済・社会的効果分析を専門的に行う機関による調査といった別の方法がとられる。

3　Kostoff（1997）は質の高いピアレビューの一般原則について，以下の項目を挙げて議論を行っている。ピアレビューを行う機関のシニアマネージャーの支援，ピア・レビューの運営者やリーダーのモチベーション，評価者・評価グループの能力と客観性・信頼性の確保，パネル間や分野間での標準化，評価基準の選択，評価者・被評価者双方の秘匿性および評価者の責任意識，コスト，倫理基準の保持。

の中の一つとして，評価者が評価対象側から提出された研究成果だけを資料として判断を下すのではなく，様々な定量的・定性的情報をも参照して多様な情報のもとで評価を行う必要性が指摘される[4]。定量的情報の中でも学問的な質が評価基準である場合には，「ビブリオメトリクスBibliometrics」（計量書誌学，書誌計量学，文献計量学などと訳される）がその一つの例として挙げられる。ビブリオメトリクスは一般的に，論文数や引用数といった数量の分析や，直接的な引用関係および専門用語や被引用文献の共出現関係の分析を基にして，研究活動の特徴を定量的に示そうとするものである[5]。この中でも，論文数は研究活動の生産性の高さを示し，引用数はその論文がその後の研究へ与えたインパクトの大きさを示すと考えられるため，欧米諸国では1980年代から研究評価に頻繁に用いられてきた[6]。しかし日本ではこれまでビブリオメトリクスの研究評価への利用は極めて限定されたものであり，それゆえに評価における定量的情報の欠如に対する批判がしばしばなされている（例えばSwinbanks, Nathan and Triendl 1997）。他方で，このようなビブリオメトリクス指標が持つ意味は不明であり問題点も多いため評価に使うべきでないという拒絶的な反応が示されることもある。さらには，一部の大学ランキングなどではビブリオメトリクスの技術的な問題を考慮せずに安易な形で分析を行い，その結果のみが一人歩きしてしまうこともある。

このような日本の現状を鑑み，本稿ではビブリオメトリクスが日本という非英語圏の国でどの程度，学問的な質の評価におけるピアレビューへの参照情報として利用可能であるかを検討する。事例として大学評価・学位授与機構（NIAD）が2000～01年度に行った理学分野の大学の研究評価を取り上げる。

２．ビブリオメトリクスとピアレビューとの結果の整合性

ビブリオメトリクスがピアレビューを支援するツールとして有効であるとは，どのような事を意味するだろうか。一つには，ビブリオメトリクスによってピアレビューの結果とある程度同等のものを，ピアレビューよりも少ないコストで提示できることが挙げられる。その場合にはビブリオメトリクス指標を参照情報の一つとすることにより，評価者は全ての論文を一から精査するのではなく，例えば引用数の高い論文を選んで精読するなど，その膨大な労力を軽減することができる。もう一つは，逆にピアレビューに大きな誤判断がある場合にはビブリオメトリクスがそれを回避するための情報を提供できることが求められる。ピアレビューでは原則的には評価者は自らと同一分野の研究のみを評価する。しかし，大学評価のように評価対象が多様な分野の研究者を含む場合には，評価をするのに十分な専門知識を有していない研究テーマの評価もする必要が生じる可能性はある。そのような場合に，大きな誤判断を防ぐための情報を提供できることがビブリオメトリクス指標に求められるであろう。

この両面を検討するには，ビブリオメトリクスとピアレビューの両者の結果がどの程度一致し，また，どこで両者の違いが生じているかを分析する必要がある。Narin et al.（1976）に

---

[4] Braun（1999）は，これはあたかも医者が診断を行う際に，問診という定性的な方法だけでなく，血液や尿の定量的な分析結果をも参照しながら最終的に判断を下すのと同様であると述べている。

[5] ビブリオメトリクスの研究は科学技術政策研究やリサーチ・オン・リサーチ（研究活動の研究）からと，図書館情報学からの両面から行われている。ここで述べた定義は主に前者で行われている研究に相当し，「サイエントメトリクスscientometrics」（科学計量学）と呼ばれる分野の中心を占めるものである。なお，図書館情報学においては，購読ジャーナルの選定や文献検索システムにおける類似文献の識別など，別の目的で行われている研究も多い。

[6] 当時のレビューとしてはGibbons and Georghiou（1987）, Office of Technology Assessment（1986）を参照。

よればビブリオメトリクスとピアレビューの結果の整合性の検証は，既に1960年代から焦点を置かれてきた。ピアレビューの結果は詳細には公表されないことが多いため，先行研究の多くでは公表されている大学ランキングや大学評価の結果との整合性の検証が行われている[7]。米国ではCartter （1966）やRoose and Andersen （1970）による教員への評判調査の結果との比較が行われており（例えばAnderson, Narin and McAlister 1978），英国では資金配分を行うための大学評価であるResearch Assessment Exerciseの結果との整合性が検証されている（Irvin 1989, Zhu, Meadows and Mason 1991, Oppenheim 1995, 1997）。オランダでは大学協会（VSNU）が行う大学評価の幾つかの分野について，ライデン大学科学技術論センター（CWTS）にビブリオメトリクス分析を委託しており，その結果が評価者に提示されるとともに，整合性の分析もなされている （Rinia et al. 1998, 2001）。これら先行研究はいずれも両者の間で整合性が高いことを認めており，研究評価のための一つの情報としてビブリオメトリクスを利用することを正当化する根拠となってきた。だが，これらは大学という機関全体レベル，あるいは，その内部の研究グループを対象としたものである。一般的に，機関全体レベルでは規模が大きければ著名な研究者の数も多くなり，機関全体の論文数も多くなるために，ピアによる機関の判断とビブリオメトリクス指標との整合性が高くなることは比較的容易に期待できる。他方で，組織を構成する教員レベルでの評価では，どの程度ビブリオメトリクス指標がピアレビューと整合性を持ちうるかは明らかではない。そのため，本論では教員個人レベルでの整合性に焦点を置いた分析を行う。

３．分析方法

3.1 事例分析対象

ビブリオメトリクスに限らず特定の評価手法を用いる際には，評価対象の規模，評価を行う単位，評価項目などにあわせて，テーラーメイドに方法の設定・修正を行う必要がある。そのため，まず分析対象を設定する。本分析では NIAD が2000～01年度に行った理学分野の研究評価を対象とする。この評価で対象となったのは５つの大学の理学部と1つの大学共同利用機関である。評価は基本的には学部・研究科を評価の単位として研究体制・支援体制や方策などを評価するものであるが，研究成果については各組織を構成する教員ごとに業績の「判定」を行い，それを積み上げることで組織レベルの評価とする方法をとった。

教員個人レベルの「判定」では，理学分野をさらに，数理・情報科学，物理学，化学，生物科学，地球科学，天文・宇宙科学の6領域に区分し，各教員はここから１つ，ないし複数の領域を選択して，最近5年間（1996年～2001年）の研究活動の概要や研究業績一覧，および代表業績そのもの５編以内を提出することが求められた。各領域では，その分野の専門家である「評価員」から構成される「部会」を作り，提出された情報を基に２名以上の評価員が各教員の研究業績の「判定」を行った[8]。判定では研究の学問的な質（「研究内容及び水準」）と社会的効果（「社会（社会・経済・文化）的貢献」）の２項目が個別に評価され，前者の研究

---

[7] 大学評価以外を対象とするビブリオメトリクスとの整合性の分析（例えば，ジャーナルの評判との整合性，教員の評判や受賞歴との整合性，競争的資金採択可否との整合性）も行われている。それら先行研究についてもNarin et al.（1976）を参照。

[8] 評価の際に評価員にはジャーナルのインパクトファクター（IF）の一覧を参考資料の一つとして配布している。また，各評価員が独自に*SCI*を検索することは妨げていない。そのため，評価員の評価がビブリオメトリクス指標と全く独立に行われたとは言えない。

水準については判定は４段階（「卓越」「優秀」「普通」「要努力」），および研究評価の対象には当たらない「該当せず」に区分された。この4段階の定義は各領域ごとに文章で設定されて公表されている。個々の評価員は各教員についての一次的な判定を行い，その結果を各部会に持ち寄り，討議を経て最終的な判定結果を確定するというプロセスがとられた。教員ごとの最終的な判定結果は，各機関の各領域ごとで集計され，その集計値が報告書に記載され公表されている[9]。本分析では，研究水準について，集計を行う前の各教員ごとの最終判定結果をピアレビューの結果として用いる。

3.2 ビブリオメトリクスの利用における技術的改善

上記の「研究水準の判定」におけるピアレビュー結果との整合性を分析するためには，ビブリオメトリクスの使用方法も若干，精緻化する必要がある。これまでビブリオメトリクス自体にも様々な方法論上の欠点があることは指摘されている（例えば Martin and Irvin 1983, Schubert 1996）。一つは分野ごとに指標の平均値が異なることである。論文数については論文を産出しやすい分野とそうでない分野があり，論文の引用数についても平均値は分野によって大きく異なる。そのため，ピアレビューと同様にビブリオメトリクスでも分野を超えた比較は困難とされる。また別の問題として，データベース自体に収録されているジャーナルの分野，出版国，言語の偏りがあり，特に非英語圏である日本では研究成果のどの程度が分析できるかは不明である。また，データベースには入力誤りや表記揺れなどの問題もある。さらには，引用数の計測において，自己引用（自分で自分の論文を引用すること）を過度に行って引用数の「水増し」を行っている場合をどう扱うかも問題となる。そこで，本分析では評価の実務作業に定常的に利用できる範囲で次のような改善を行った。

(1)　被引用数の分野間標準化

今回実施された理学系の研究評価では，各教員は提出する「業績一覧」に必ずしも全ての業績を記入することは求められず，何をどの程度記載するかは教員側に委ねられた。そのため教員によっては主要な業績のみを記入した人もいれば，学会発表も含めて網羅的に記入した人もいた。このように全ての業績を網羅的に記入することを求めなかった理由は，研究を単純にアウトプットの量だけで判断するのではなく，その内容の質に焦点をおいて判定を行うとしたためである。また，そのような理由から評価員は提出された５編以内の代表業績そのものに実際に目を通して判定作業を行った。このことから，ビブリオメトリクス分析においても単純に論文数の多さを比較することは適さず，より研究の質の側面を反映していると考えられる引用数を中心に分析を行うことが妥当であると考えられる。また，そのために引用データを含むデータベースである ISI 社の *Science Citation Index （SCI）* を用いる。*SCI* はビブリオメトリクス分析に最も用いられるデータベースであり，自然科学分野全般を対象としてデータを収録しているために，本分析対象のように，分野が多岐に渡る際には適している。

前述のように引用数の計測においても平均値は分野によって異なる。それは６つの領域の間だけでなく，領域の内部でも異なる。例えば，「物理学」領域でも素粒子物理学と物性物理学では引用数の平均値は一般的に数倍異なる[10]。そのため，領域内部においても分野間の標

---

[9] 各教員個人の判定の結果は非公開である。

準化の方法を考える必要がある。分野間での標準化の方法はこれまでに幾つか提案されている（Shubert and Braun 1996）。標準化ではある論文群を参照対象として設定し，その中での平均値との比や位置づけ（順位）を指標とする方法が用いられることが多い。参照対象の設定には，1）その論文が掲載されたジャーナルの全論文を用いる方法，2）データベースにより付与されたその論文の学問分野分類に属する全論文を用いる方法，3）共引用分析や共語分析などの方法により類似性の高い論文を独自に設定して参照対象とする方法がある。この内で1）については，*Science*や*Nature*などの引用の平均値が高いジャーナルに掲載された論文は，たとえ多く引用されても平均値に達しない場合には過小評価されてしまう問題がある。3）については，閾値の設定によって分野の大きさは異なり，複数の分野間の比較に適した一意の閾値の設定の仕方は明らかでない。そのため，本分析では2）の当該論文と同じ学問分野に属する全論文を参照対象として用いる。分析に用いた*SCI*では，分野カテゴリ（Subject Category）が170程度設定されており（年によって若干異なる），各ジャーナルごとに１つ以上の分野カテゴリが付与されている。そのため，分析対象の論文が掲載されたジャーナルと同じ分野カテゴリが付与されたジャーナルに掲載されている全論文を参照対象とする。なお，分野カテゴリが複数付されている場合には論文数は分数カウントし，それぞれの分野カテゴリに割り振る。そのため，複数の分野カテゴリが付与された論文については，一つでも同じ分野カテゴリを有するジャーナルの全論文が参照対象となるが，分数カウントを行っているため分野カテゴリの重なりが大きい論文ほど1を最大値として加重カウントされることになる。具体的な分析作業においては，*SCI*（CD-ROM版）に毎年約80万件収録されている全論文（Article, Review, Letterのみを対象とし，Meeting-abstractやEditorial letterなどは含めない）について，2001年までの引用回数をそれぞれ計測し，分野カテゴリごとにその平均値と分布を算出して標準化を行う。ただし，*Science*や*Nature*などの「学際分野」の分野カテゴリが付されたジャーナルでは，実際には物理学や生物学などの様々な分野の論文が掲載されている。そのため，個々の論文について，その論文が引用している論文のジャーナルの分野カテゴリを集計し，上位３つの分野カテゴリを，個別の論文の分野カテゴリとして用いる。

また，指標化の方法としては，この参照対象の中で当該論文が引用数による序列で上位何%に入るかという順位を指標とする方法を用いる[11]。ただし，数年前に出た論文は引用される期間が十分でなく，引用数0回となるものも多いことを鑑み，各論文でなくその論文が掲載されたジャーナルの平均引用数の分野カテゴリ平均との比を指標とする場合も試みる[12]。ここでは1996年の論文が2001年までに引用された回数のジャーナル内での平均値を分野カテゴリ全体の平均値と比した値，および1997年の論文についての同様の比の値を求め，その２年間

---

[10] *SCI*の分野カテゴリで「素粒子物理（UP）」が付与されたジャーナルの全論文と，「物性物理（UK）」が付与されたジャーナルの全論文で，2000年に出版された論文が2002年までに引用された回数の平均値を求めると，前者は4.5回，後者は1.9回であった（計測には*SCI*のCD-ROM版を用い，Article, Letter, Reviewのみを対象とした。分野カテゴリが複数付与されているときは分数カウントを行い，自己引用は除外した）。

[11] 引用数の序列（順位）を指標に用いた理由は，教員の業績の判定作業でも4段階に序列化を行っていることと適合すると考えられるためである。ただし，分野カテゴリ全体の平均値との比を指標とする方法も試行的に行ったが，いずれの領域においても，順位を用いる方法との間で相関係数に大きな差は生じなかった。

[12] 各ジャーナルに掲載された論文の平均引用数を示す指標として「インパクト・ファクター（IF）」が広く知られている。概して言えば，IFは各ジャーナルの論文の出版後１～２年間の平均引用数を示すものである。それに対して，ここでは出版後４～５年間というより広い期間の平均引用数を基にし，さらに，その値と分野カテゴリ内の全論文の平均引用数との比をとることにより標準化した指標を用いる点で異なっている。

の平均値を用いる[13]。

(2) 出版年，文書形式の区分

参照対象とした論文は，分野カテゴリが同一であるのに加え，出版年が同一のもののみとする。これは出版年が古い論文ほど引用される期間が長くなるため，出版年で区分しないと古い論文ほど有利になるためである。また文書形式も Article, Review, Letter とあり，それぞれで引用の平均値に違いがあることも知られている（Amin and Mabe 2002）。そのため，同一の文書形式のみを参照対象とする場合と，区分せずに全てを参照対象とする場合の双方を試行して比較を行う。

(3) 自己引用の除去

同一著者名を含む論文からの引用を自己引用と機械的に推定し，*SCI* 上の全論文について自己引用を除外して計測する場合としない場合の双方の引用数を分析して比較を行う。*SCI* においては著者のファーストネームはイニシャルのみの表記であるので同姓同名の別人が混在している可能性はあるが，同姓同名の別人からの引用が引用全体の中で占める比率は大きくないと予想されるため，この方法を用いる。

(4) その他の補正

著者名のミドルネーム有無の表記揺れ，引用文献の記述の仕方の表記揺れなどは機械的に補正する。

このような方法を用いることで，引用数の指標についていくつかの指標化の選択肢により分析を行う。さらに指標化する分析対象として，実際に提出された代表業績5編を用いる場合と，それに拠らずに業績一覧に記された論文の中で引用の指標が高い５編を用いる場合の両者を検討する。ただし，提出された代表業績の中で *SCI* で検索された論文数が５編に満たないときは，業績一覧から引用の指標が高い論文を選択して合わせて５編とする。このようにして得られた５編分の指標を集計して５で除算することにより，各教員の指標の値とする。なお，業績一覧に記された全業績の中で *SCI* で検索されたものが５編に満たない場合にも，単純に5で除算する。これにより，５編に満たない数の論文しか *SCI* に収録されていない場合には，評価対象教員の指標の値は低くなる。これは今回の評価対象である理学系の多くの分野では５年間で５編以上の論文を産出することが標準的であると予想したためである。

４．結　果

4.1 研究アウトプットの *SCI* 収録割合

分析に用いたデータベースである *Science Citation Index* に収録されているデータのほとんどはジャーナル（学術雑誌）に掲載された論文である。そのため，書籍や報告書などのその他の形態のアウトプットについては分析を行うことはできない[14]。さらに，*SCI* に収録されて

[13] ただし，1997年以降に創刊された，あるいは *SCI* に収録されるようになったジャーナルは1997年以降のもっとも古い年２年分を用いている。

[14] 引用の分析に関しては，それ自体が *SCI* に収録されていないジャーナルや書籍や報告書であっても *SCI* に収録されているジャーナルの論文に引用されている回数を測定することはできる。しかし，通常，論文は同じジャーナルの論文に引用されることが多いことから，このような測定値は過小評価となる可能性がある。そのため，今回は分析を行わなかった。

いるジャーナルはほとんどが英文誌であり，日本語の論文を多く産出する分野では分析には適さない。そのため，まずは提出された業績のどの程度の割合を本ビブリオメトリクス分析で分析可能であるかを明らかにする必要がある。

表１は各部会に教員から提出された「代表業績５編以内」の中で「英文のジャーナル論文」の割合(a)，さらにその中で*SCI*に収録されていた割合(b)を示している[15]。なお，日本で出版されているジャーナルや紀要であっても，英文誌である場合は英文ジャーナルとしてカウントした。表からは物理，化学，生物科学では代表業績として教員が提出したものの９割程度が英文のジャーナル論文であったことが分かる。数理・情報科学でも８割程度は英文論文が提出された。他方で，地球科学や天文・宇宙科学では英文ジャーナル論文は６割程度である。地球科学では日本語の論文や報告書が多い。天文・宇宙科学領域においては，プロシーディングスが多いのに加え，巨大装置を用いた研究を行っている機関からは装置の仕様・設計図や観測・運転記録なども提出されており，評価対象機関の特徴も反映されていると言える[16]。次にそれら英文ジャーナル論文のうちで*SCI*に収録されていた割合も，やはり物理，化学，生物科学で９割程度と高かった。地球科学，宇宙・天文科学においても英文ジャーナル論文の７～８割は*SCI*に収録されていた。それに対して数理・情報科学は５割以下と収録率が低い。これは数理・情報科学領域においては英文の大学紀要に掲載された論文が多く，*SCI*にはほとんど収録されていないことが影響している。この結果，数理・情報科学，地球科学，天文・宇宙科学では実際に提出された代表業績５編以内の内の半分以下しかビブリオメトリクス分析の対象となっていないことになる。

表１　部会ごとの *SCI* での検索ヒット率

部　会	研究者数	「業績５編以内」			「業績一覧」全体の内訳			
		英文ジャーナル論文の割合 (a)	左の中でSCIに収録されていた割合 (b)	(a×b)	英文ジャーナル論文の割合 (c)	左の中でSCIに収録されていた割合 (d)	(c×d)	*SCI*で検索された全論文数 (e)
数理・情報	127	81.3%	47.7%	38.8%	56.8%	40.4%	22.9%	280
物　理	170	90.2%	87.3%	78.7%	84.9%	87.3%	74.1%	2,316
化　学	146	93.7%	93.3%	87.4%	85.2%	91.7%	78.1%	2,358
生物科学	132	88.5%	87.4%	77.4%	78.1%	79.6%	62.1%	1,073
地球科学	128	62.7%	71.8%	45.0%	48.1%	59.9%	28.8%	560
天文・宇宙	183	62.8%	78.7%	49.4%	65.4%	83.3%	54.5%	1,335

表１ではさらに，実際に提出された５編以内だけでなく，業績一覧に挙げられた全ての業績を対象とした場合の結果も右に示している。ただし，先述のように「業績一覧」は必ずしも教員の５年間の全ての業績が網羅されているものではないことには注意が必要である。表からは，代表業績５編に加えて記入されているアウトプットには英文ジャーナル論文以外の報告書やプロシーディングなどが入り，英文ジャーナル論文の比率は天文・宇宙科学を除き

[15] 5編より少ない数の資料のみを提出した教員も多く存在した。また，5編以内という制限にも関わらず，間違ってそれ以上の数の資料を提出した教員も少数存在した。表は，それらの修正は行わず単純に提出された資料の数を計算したものである。

[16] この値は分野や評価対象機関の特徴に影響されているだけでなく，各機関の資料の提出の仕方にも影響を受けている。特に今回の評価は初めての試みであったために，どのような資料を提出すべきかのコンセンサスが明確でなく，機関によっては論文などの業績そのものではなく，研究活動全般が分かる資料を提出している教員もいた。そのため，評価対象者が少数の機関に偏っている分野ではこれらの影響は無視できないと思われる。

下がっている。それでも，数理・情報や地球科学ではビブリオメトリクス分析の対象範囲の割合が物理，化学，生物科学と比べて低いことには変わりはない。

これらの結果から，物理，化学，生物科学についてはピアレビューにおいて評価員が評価したアウトプットとほぼ同じものをビブリオメトリクス分析においても分析可能であるが，その他の分野では半数程度しか同一のものを分析できないことになる。このような制限はビブリオメトリクス分析の明確な限界であり，解釈をする上で前提とすべきである。

4.2 ピアレビュー結果との整合性

上述の多種の選択肢による測定について，ピアレビュー結果との整合性を Spearman の順位相関により測定した。表２は相関係数を示している。いくつかの選択枝で分析を行ったが，概して選択枝の間では大きな相関係数の違いは生じなかった。数理・情報科学を除き，どの選択肢においても1%有意でピアレビューとの相関関係が見られた。その中でも，物理，化学，生物科学において他分野と比べて高い相関が得られた。数理・情報科学については，論文の引用数に基づく指標では有意な相関関係は見られなかったが，ジャーナルの平均引用数を基にした指標ではピアレビューとの相関が見られた。

各選択肢について，その他の条件を同じくしたもので比較した場合，分野によって結果は多様ではあった。文書形式については，ほとんどの場合で区別する場合と同一に扱う場合とで小数点２桁まで相関係数に違いは生じなかったため，表２は統一的に区分しない場合のみの値を示している。文書形式で違いが生じなかった理由は，大半の論文が article であり review や letter の割合自体が圧倒的に少なかったことが挙げられる。

表２　ピアレビュー結果との順位相関

指標化の選択肢			部会					
対象	指標の基礎	自己引用	数理・情報	物理	化学	生物科学	地球科学	天文・宇宙
提出された業績５編	論文引用数の分野内順位	含む	0.19	0.56**	0.58**	0.60**	0.35**	0.38**
		含まない	0.16	0.49**	0.52**	0.57**	0.37**	0.37**
	ジャーナル平均の分野平均比	含む	0.30**	0.44**	0.66**	0.57**	0.41**	0.30**
		含まない	**0.31****	0.42**	0.66**	0.56**	**0.43****	0.29**
業績一覧の中から5編	論文引用数の分野内順位	含む	0.19	**0.59****	0.61**	**0.66****	0.38**	**0.42****
		含まない	0.15	0.55**	0.62**	**0.66****	0.40**	0.41**
	ジャーナル平均の分野平均比	含む	0.30**	0.48**	**0.68****	0.62**	0.41**	0.31**
		含まない	0.30**	0.47**	0.68**	0.62**	0.43**	0.30**
論文などの数			0.47**	0.49**	0.57**	0.61**	0.36**	0.47**

* P<0.05　** P<0.01

また，９割以上で，業績一覧から引用指標の高い５編を選んだ場合の方が，実際に提出された代表業績５編に限定する場合よりも相関係数は高かった。また，７割以上で，自己引用を含む場合の方が除去する場合よりも相関係数が高かった。業績一覧から５編を選択する場合が相関が高かった理由は明確ではないが，ピアレビューでも評価員は「業績一覧」に挙げられた全論文のタイトルやジャーナル名も見て総合的に評点を付けていることとは適合している。また，自己引用を含む場合の方が相関が高い場合が多かった理由は，当該研究者が論文を多く産出していなければ自己引用の数は多くならならず，そのような生産性の高さがピ

アレビュー結果とも適合した可能性が挙げられる。また，出版されてからの引用期間が短いために，自己引用を除去すると引用数０回の論文が多くなり，差違化ができないことも挙げられる。

数理・情報科学では論文の引用数を用いた指標との相関は有意でなかったが，ジャーナルの平均引用数の分野全体との比を用いた指標では有意な結果が出た。だが，他の分野においても２つの指標化の方法で相関係数に大きな違いはなかった。数理・情報科学だけでなく地球科学および化学においても，ジャーナルの平均数平均を基にした指標の方が相関係数が若干高い値となった。本来，ジャーナルの平均引用数は，必ずしもその論文の引用回数が高いことを保証するものではない。だが，平均引用数の高いジャーナルでは，論文の掲載可否の審査において厳しい評価がなされており，論文が掲載されたこと自体が既にレフェリーからその質を保証されたことを示す。そのため，ピアレビューにおいてもジャーナル名は判定を行うための有効な情報となったことは通常考えられる。さらに，本 NIAD の評価のように出版から数年しかたっていない論文も多い場合には，引用期間が短く引用数が０回のものも多くなる。実際，引用があまり行われにくい数理・情報科学では，論文の引用数に基づく指標では38%の教員が５編の引用数の合計が０となった。このため，ジャーナルの平均引用数が相関が高くなったと考えられる。

表２には引用数による指標だけでなく業績一覧に示された論文等の数との spearman 順位相関係数も示している。ここでは，英文および邦文の論文の合計数，それにプロシーディングの数を足した数，さらに書籍やその中の章の数を足した数の３種類の内で最も相関が高かったものを示している。ただし先述のように，業績一覧には必ずしも５年間の全ての業績が記入されていないので，一つの参考指標として扱うべきものである。その結果では，数理・情報科学および天文・宇宙科学では論文数との相関が引用数の相関よりも高く，他分野でも引用数の相関と大きな差違は出ていない。そもそも引用数の指標と論文数との相関自体も高い。表２の最下部には引用数（業績一覧の５編を自己引用を含まずに順位により指標化したもの）と論文数（英文・邦文論文とプロシーディングの総数）の spearman 順位相関係数も示している。この結果からは，論文生産性の高い研究者は被引用数の高い論文を産出しているという傾向が全体的には存在することを示している。

このように引用および論文数について様々に指標を設定したが，実際には概して，それらの間で大きな差は出ず，ほとんど全てが有意な相関を示している。このような傾向は既に他国の整合性の先行研究でも見られていることであり，Martin と Irvin らはこのように一つ一つの指標はその意味や有効性は必ずしも明確でなくても，複数の指標（本来はビブリオメトリクス指標のみならず，受賞や招待講演数などの様々な指標）がある程度同一の結果を示し，またピアレビュー結果も同様の結果を示す場合には，その評価結果（ピアレビューおよび各指標）の信頼性が高いことを期待できるという考えを示している （Martin and Irvine 1983, Martin 1996)。

表３では，論文の引用数による指標を基にして，ピアレビュー結果と同様の割合で「卓越」から「要努力」までの４段階の評点に区分した場合に，ピアレビューとどの程度異なっていたかを示している。引用数による指標は,「代表業績に拠らずに５編を, 自己引用を含む形で, 引用数の順位で指標化する方法」に統一しており，例えばピアレビューで「卓越」となった割合と同じ割合だけ，指標が上位の教員を「卓越」とした。いる。その結果，概して，全部

会で60%程度がピアレビューと同じ評点であり，１段階異なるのは40%程度であった[17]。２段階以上の差がついたものは数%にすぎなかった。参考のために表３の右には，ピアレビューと同様の割合にランダムに評点付けした場合に生じる，ピアレビューとの差異（期待値）を示している。それら値と比較しても，特に２段階以上の差がつく場合は少ないと言える[18]。

表3　ピアレビューとビブリオメトリクス分析との相違

部　　会	同　　じ	１段階異なる	2段階以上異なる
数理・情報	52%	45%	3%
物　　理	55%	43%	2%
化　　学	57%	38%	4%
生物科学	60%	37%	3%
地球科学	65%	27%	8%
天　　文	59%	41%	1%

（参考）同割合にランダムに評点付けした場合に生じる差異（期待値）		
同　じ	１段階異なる	2段階以上異なる
44%	46%	11%
37%	48%	15%
35%	45%	20%
34%	48%	19%
43%	44%	13%
44%	47%	8%

これらの値は，実際には，ピアレビューの過程において，同一の教員に対して二人以上の評価者が最初につけた判断（部会で最終的な合議を行う前の一次的判断）で差違が生じた割合とほとんど変わらないものであった（一次的判断では，一段階の差異は全分野平均で35%生じており，２段階以上も３%程度生じた)。NIADの判定作業ではこれら一次的結果を持ち寄って議論し，最終的な判定結果を確定するというプロセスをさらに経たが，議論の仕方によっては１段階の変動は生じる可能性があるものと言える[19]。ピアレビューやビブリオメトリクスの方法を様々に改善したとしても常に同じ結果を生むような完全な方法を期待することは通常，困難である。この事を前提とすれば，評価結果の利用方法の側を，１段階の差違で全体的な評価結果やそれによる影響が大きくならないように設定すべきである。他方，２段階以上という大きな差違はピアレビューの中でも，ビブリオメトリクスとの間でも，頻繁には生じなかった。そのため，そのような結果が出た場合にはその判断の妥当性を再確認することが必要とされる。ビブリオメトリクスはこのように，ピアレビューの結果の再検証のための参照情報として利用できることが期待される。

4.3　評価結果の相違の原因

(1)　評価員のコメント

では，ピアレビュー結果とビブリオメトリクス分析の結果との差異はどのような場合に生じていたのであろうか。表４はピアレビューがビブリオメトリクス分析よりも２段階低く評価している場合に，評価者が作業シートに付したコメントをまとめたものである。コメント

---

17 オランダのライデン大学科学技術論センターのvanRann教授へのヒアリングによれば，同センターがオランダ大学協会の委託で行った生物学の研究グループ単位の分析では，大まかに言って25%でピアレビューとビブリオメトリクス分析の結果が異なっていたと言う。ただし，評価自体の方法やビブリオメトリクスの分析方法が本分析とは異なるものであるため単純に比較はできない。

18 ビブリオメトリクス分析の結果を知ることで，ピアレビューの評点付けがいかに容易になるかは，情報量として計算可能である。すなわち，何の情報も無く４段階に当該割合で区分する不確実さをH（x）とし，ビブリオメトリクスの結果を知った上で４段階に当該割合で区分する不確実さをH（x|y）とすれば，相互情報量は不確実さの減少としてI（x;y）=H（x）-H（x|y）で表せる。また，I（x;y）/H（x）により不確実さが減少した割合が示せる。各領域についてI（x;y）/H（x）を計算すると，数理・情報科学9.4%，物理学18.5%，化学19.4%，生物科学25.3%，地球科学18.6%，天文・宇宙科学14.0%となった。

19 ピアレビューにおいて一次的な結果で差異が出た原因の最大のものは，どの程度の研究内容を４段階の各評点に対応させるかという水準の定義が十分詳細には明文化されていなかったことが挙げられる。そのため，相対的に厳しめの評点をつける評価員と，易しめの評点をつける評価員が生じることになった（この問題については5章で述べる)。なお，次年度以降のNIADの評価では，事前に水準の定義を詳細に設定するよう変更している。

には，提出書類に研究内容の説明が記入されるべきにもかかわれず，それがないことや，当該教員の関与の割合が不明などの説明不足を指摘するもの，「筆頭著者論文がない」「単著が少ない」ことを厳しく評価しているものが見られた。共著については，ビブリオメトリクス分析でも著者数で分数カウントすることは機械的に可能であるが，関与の割合とは必ずしも等しくないため行わなかった。これが差違を生じる要因の一つになったことも予想される。また，特に否定的なコメントは記入されていない場合もいくつかあり，ビブリオメトリクス分析では判断できない内容があったのかは確認できない。

表4　ピアレビューの方が2段階低く評価したものに付されていたコメント

コメント	件　数
筆頭論文や単著がない	4
論文数や研究のレベルが低い	3
研究内容の説明がない，不十分。	2
本人の関与割外が不明	2
（その他，特に否定的なコメントなし）	6

表5　ピアレビューの方が２段階高く評価したものに付されていたコメント

コメント	件　数
新規分野・新規理論への挑戦，独創性高い	5
世界をリード，日本のトップレベル	2
特殊で貴重な分野で研究を行っている	1
（その他，特に肯定的なコメントなし）	24

他方で，２段階高く評価しているケースでは，新規分野や新規理論を開拓している，世界をリードしているなどのコメントが見られた。特に新規で独創的な研究は引用数が短期的には高くならない可能性があるため，ビブリオメトリクス分析では現れないインパクトをピアレビューで高く評価した可能性がうかがえる。

⑵　評価員による専門分野の包含との関係

一般的にピアレビューの信頼性が脆弱になる要因の一つとして，評価対象が評価員の専門分野によってカバーされていない場合が挙げられる。特に，ジャーナルへの投稿論文のピアレビューや，特定領域の研究プロジェクト選定のためのピアレビューと異なり，大学評価では多様な研究分野の教員が一つの学部や学科に属している可能性があり，限られた数の評価員ではその分野がカバーされない恐れがある。そのため，NIAD の評価においても評価員の専門分野を外れる分野の研究を行っている教員の判定について，ピアレビューの結果とビブリオメトリクスの結果が大きく異なるか検討した。先述のように教員が産出した各論文は，*SCI* に収録されている場合に限って，ジャーナルごとに分野カテゴリーが設定されている。評価員についても同様に，過去５年間の論文をその氏名と所属機関名（過去５年間に転籍している場合は，以前の所属機関名も含む）を用いて検索を行い，それら論文の分野カテゴリを同定した。

その結果，どの評価員も１本も論文を産出していない分野カテゴリーに，半分以上の論文を産出している教員は６領域全体で13名のみであった（論文が *SCI* に1本しか収録されていない教員３名を含む）。それらについて，ピアレビューとビブリオメトリクスの評点の違いは表６のようになった。結果的には，判定全体と比べて顕著な差違は生じなかった。

表６　評価員により分野がカバーされていない教員のピアレビューとビブリオメトリクス結果の差違

	同じ	１段階異なる			２段階異なる
			高い	低い	
教員数 （割合）	8 (61.5%)	5 (38.5%)	2 (15.4%)	3 (23.1%)	0 (0%)

(3) 職位との関係

また，ピアレビュー結果に主観性が入る要因の一つとして，若い研究者への過小評価や，著名な教員の過大評価がしばしば挙げられる。NIAD の評価では年齢や教員の一般的評判は資料として存在しないため，これらの変数と関係が深いと考えられる職位（教授，助教授，講師，助手）について，ピアレビューとビブリオメトリクス分析との差違に違いがあるかを分析した。その結果では，６領域全体で見ると，教授の方がピアレビュー結果がビブリオメトリクスの結果よりも評点が高くなり，逆に講師・助手についてはピアレビュー結果が低いという傾向が若干見られた。独立性の検定では，ピアレビューとの差違と職位の２変数は５％有意で独立ではなかった。ただし，各領域ごとに分けた分析では，１つの領域のみで５％有意となり，その他の領域については統計的に有意な関係は示されなかった。

表７　ピアレビューとビブリオメトリクスとの差違と職位との関係

	総　数	ピアレビュー結果の方が評点が高い	同　じ	ピアレビュー結果の方が評点が低い
全　　体	776	20.5%	57.9%	21.6%
教　　授	276	24.6%	58.7%	16.7%
助　教　授	229	18.8%	59.0%	22.3%
講 師 ・ 助 手	271	17.7%	56.1%	26.2%

χ^2=9.7　P<0.05

５．評価員間・部会間での評点基準の調整の支援

ここまで述べたように，ビブリオメトリクス分析はピアレビュー結果とある程度の相関は示したが，40%程度の割合で1段階の差異は生じており，それは評価員による評価対象の専門分野のカバーの不足や年齢による偏向などの体系的な因果関係が存在するとは言い切れないものであった。他方で，ピアレビューにおいても最終合議を得る前の一次段階においては判定結果に１段階の差違は頻繁に生じた。通常，評価者がたとえ意識的・無意識的なえこひいきなどを行わなくとも，同一の評価対象に対してピアレビューの評価者の間で評点が異なることになる要因としては原理的には次の３つが考えられる。一つはそもそも「優れた研究」とは何かという基準の不統一，一つはどの程度優れた研究を各評点へ対応させるかという評点付けの水準設定の不統一，一つは評価対象そのものを理解している度合いである。これらそれぞれは，評価者の間で基準を明示化し，コンセンサスを形成することや，評価者の選択方法などで改善できるが，ビブリオメトリクスは，引用数を用いて２つ目の評点基準の問題を支援するための標準的な基準を示すことができる。分野横断的に指標を用いることが可能であれば，個別の判定結果だけでなく評価員や部会の間での評点の調整の支援も行うことが可能となる。

5.1　評価員間での評点の水準の調整

評価員の間で，各評点の基準についてのコンセンサスが不十分な場合には，たとえ各評価員が評価対象について同じ解釈をしていても，評点は異なりうる。NIAD の今回の評価では，

各部会において，各評点の基準を文章で示すこと，および，各評価員の一時的な評点結果を持ち寄って摺り合わせを行うことによってこれを解消した。しかし，ビブリオメトリクス指標を用いることで，より明示的に扱うことが可能となる。

図は６つの部会の各評価員について，評価員の間でも厳しめの評点をつける傾向があったか（横軸）および，ビブリオメトリクス分析による評点と比べて厳しめの評点をつける傾向があったか（縦軸）を示す。この図では，「卓越」から「要努力」までを順に１，２，３，４と数値に変換してある。

横軸は次の式で示される指標である。

$$s_i = -\frac{\sum_j (r_{ij} - r_j)}{n_i}$$

ここでr_{ij}　：評価者i氏が教員j氏につけた評点

r_j :　：教員j氏につけられた最終的な評点

n_i :　：評価者i氏が判定を行った教員の数

これは，評価者i氏が教員j氏につけた評点が，最終的な合議による評点よりもどの程度高かったかを，i氏が行った判定全てについて平均したものである。感覚的に分かりやすいように負の符号をかけてs_iの値が低い方が厳しめの評価をつけたことを示すように変換している。

同様に縦軸のビブリオメトリクス指標に基づく評点とのずれは以下の式で表せる。

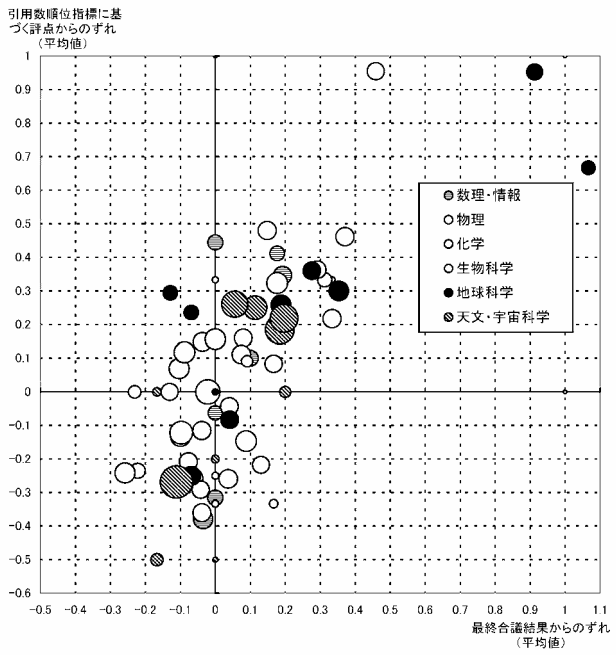

図１　各評価員の一次評価の傾向

※各円は１人の評価員を示し，円の大きさは各評価員が判定を行った教員の数を示す。
※値が低い方が厳しい評点をつけたことを意味する。

$$t_i = -\frac{\sum_k (r_{ik} - b_k)}{n'_i}$$

ここでb_j　：引用指標に基づきピアレビューと同じ割合で評点を付けた場合の教員k氏の評点

n'_i　：評価者i氏が判定を行った教員の中でビブリオメトリクス分析で対象となった（*SCI*上に少なくとも論文が1本検索された）教員の数

である。

図からは，同一の評価対象に対して，評価員の間でも厳しめの評点をつける傾向がある評価員は，ビブリオメトリクス分析の結果から見ても厳しい評点をつける傾向があることが示されている。一般的に，両指標が異なる場合としては，例えば，評価対象教員の専門分野と同じ専門分野の評価員が実際には１人しかおらず，その評価員のみがビブリオメトリクス指標からみれば妥当な評点をつけており，その他の評価員は甘い評点をつけている場合などが想定できるが，そのように両指標の値が大きく異なる評価員は存在しなかった。そのため，一次的な判定を行った後に，ビブリオメトリクス分析の結果を示して，どの程度を各評点に対応させるかを明示的に議論することにより，調整作業が容易になることが期待できる。

5.2　領域間での評点水準の調整

評価員の間での各基準の定義のずれと同様に，領域（部会）間でも大きな違いがないように調整を行う必要がある。摺り合わせが作業プロセスに組み込まれた部会内部とは異なり，部会の間では各評点の基準は大きく差が出る可能性があり，領域によっては他領域よりも大幅に甘い評点付けが行われる可能性はある。

そもそも領域間の調整の可否については２つの異なる見解がある。一つは領域ごとに評価で重視される項目は異なるため，領域を越えた調整は不可能であり，判定結果は複数の領域を通して見ることは原理的に出来ないという意見である。もう一つは，分野間での調整は広いバックグラウンドを有する共通の評価者を用いれば可能であり（Kostoff 1997），また，評価者は自己の専門分野の隣接領域もある程度は評価できるので，評価者の評価可能な領域が網羅的に重なりあうことによって科学全体の斉一的評価基準が設定可能というものである（Polanyi 1962）[20]。NIADの今回の理学系研究評価では，委員会構成は各領域ごとの部会の上位に「理学系研究評価専門委員会」が存在するため，後者のような調整を専門委員会によって行うことが可能な構造にはなっていた。だが，基本的には各領域の部会を尊重し，部会間は独立というスタンスで結果を示した。

通常，研究評価に限らず，どのような評価においても，対象は個々に多様な特徴を有するものである。評価では，評価を行う目的に即して，便宜的にある特定の項目について横断的に評価を行ったり，評価項目の重みを対象ごとに可変にしたりなどの方法をとる。そのような場合にビブリオメトリクスは横断的な指標の一つとして，複数の領域について同一方法による分析結果を提供できる。

---

[20] Polanyi（1962）は次のように述べている。「科学者はたしかに科学の小さな部分についてしか十分な判断を下せない。だが，自己の専門的研究に隣接する領域なら通常は判断できるのであり，それは他の科学者が専門とする分野を含むほど幅広いものである。それゆえ，科学者が正しい批判的判断を下すことのできる領域の間にはかなりの程度の重なり合いがある。ある重なり合いのグループのメンバーである科学者は当然，別のグループのメンバーでもあるので，科学の全体が重なり合った隣接領域からなる連鎖とネットワークによっておおわれることになろう。（中略）これら重なり合った隣接領域を通して，科学的メリットの斉一的基準が，天文学から医学に至るまで，科学の全範囲に広く行き渡るであろう。」

図２は，部会ごとに，「卓越」から「要努力」の評点がつけられた教員の引用数の順位を基にした指標がどのように分布しているかを箱ひげ図で示している。ただし，英文ジャーナル論文の数が少ない分野では5編未満となる教員が多く過小評価となるため，ここでは4編で分析を行っている。ただし，それでも数理・情報科学では69%の教員は４編未満しか*SCI*で論文が検索されなかったため，他領域よりも過小評価となり，比較することは困難である[21]。

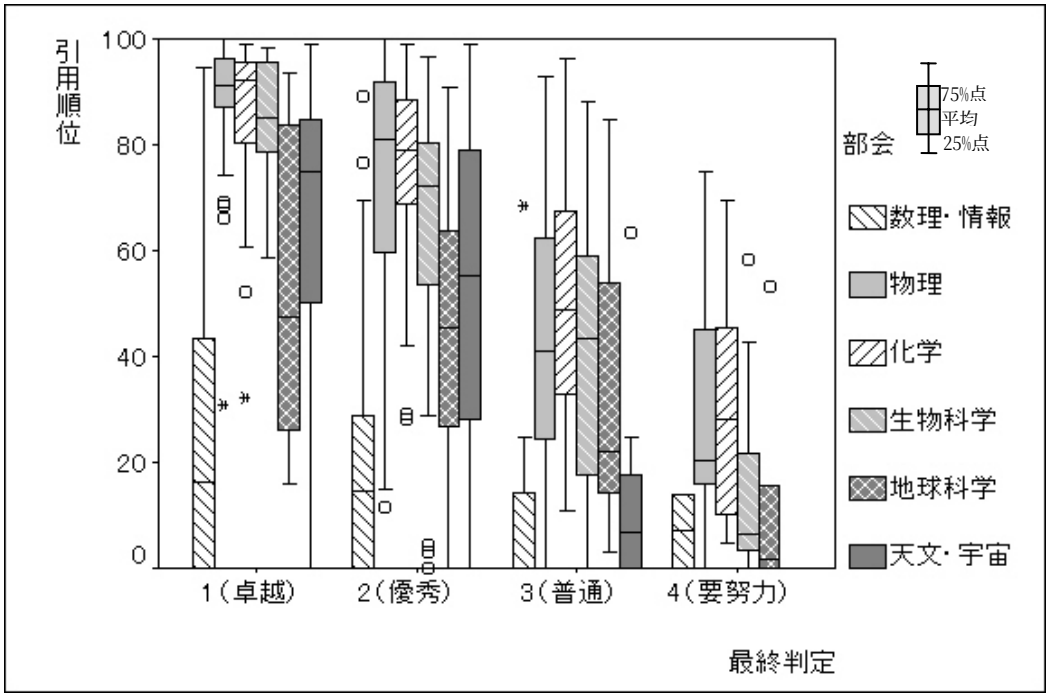

図２　各領域ごとの評価結果の基準の違い

（*SCI*に1本以上論文がある研究者のみ）

この図の結果からは，引用数という面から見れば領域間で各評点への対応付けには違いがあることが分かる。物理・化学・生物では引用数の指標で上位に「卓越」は固まっており，その平均値も部会間でほぼ近い。しかし他の領域では，これら分野よりも各評点に対応する指標の値の平均値は低く，各評点に対応する幅も大きい。最初に見たように，ビブリオメトリクス分析の対象となる範囲には分野によって限界があるため，これは必ずしも各領域の基準の厳しさを単純に示すものではないが，引用数の指標により「世界全体の論文の中で引用という形で示されるインパクトの大きさ」がいかなるものであったかという一つの基準は提供することができ，これを参照情報の一つとして調整の議論を展開することはできる。

６．結論 ～ ビブリオメトリクス手法の有効性と限界

本分析では NIAD が行った理学系研究評価を対象として，ビブリオメトリクス手法によるピアレビューの支援可能性を検討した。理学は工学や人文・社会科学と比べて国際的ジャー

[21] この値は*SCI*で1本以上の論文が検索された教員（すなわちビブリオメトリクス分析の対象とした教員）の内で，さらに1～3本のみしか検索されなかった割合である。なお，他の領域について4編未満は，7%（物理），6%（化学），28%（生物科学），40%（地球科学），22%（天文・宇宙科学）であった。

ナルへの論文投稿が多く行われており，*SCI* のような海外データベースによるビブリオメトリクス分析も行いやすいと考えられた。だが，それでも数理・情報科学や地球科学といった分野ではそれ以外の形式のアウトプットが多数産出されており，*SCI* によるビブリオメトリクス分析の対象にはならなかった。*SCI* ではなく，例えば日本の国立情報学研究所が作成している邦文論文のデータベースである「引用文献索引データベース」を使用したり，あるいは各分野に固有の海外データベースを利用すれば，いくらか対象範囲は拡大することはできる。しかし，その場合には新たに，異なる複数のデータベースによって計算された指標をいかに標準化するのかという課題が生じることになる。このように，データベースの収録範囲により分析対象が規定されてしまうことは，ビブリオメトリクス手法の大きな限界の一つであり，容易に分析可能な分野は限られることになる。だが，このことは単に分野によってデータベースが異なったり，アウトプットの言語や形態が異なることを述べるだけではない。Mode 2 と称されるような社会的な課題解決を目的とした学際的な研究活動（Gibbons et al. 1994）が増加するにつれ，論文を基にして，分野ごとに学術的な価値を評価するという方法は，そのような研究活動の評価として不十分となる可能性がある。ビブリオメトリクス分析は，評価対象の研究の目的やその活動を示すアウトプットの種類を考慮した上で，適切に利用可能な場合にのみ用いるべきである。

一方で，物理，化学，生物科学のような，*SCI* データベースで論文が十分検索できる領域では，ビブリオメトリクス分析が評価者の負担軽減のための情報を生みうることは示唆された。特に両者で 2 段階の差がつくことはまれであるため，評価者の主観的な判断の誤りにより妥当でない評価結果を生むことを抑止する情報にはなる。だが，一段階の差違は 4 割程度で起こり，またピアレビューの内部でもある程度の割合で一段階の差違は起こりうるものである。そのため，評価結果を何らかの意思決定に用いる場合には，一段階の誤差が許容できるほどの緩やかな結びつきを考えるべきであると言える。

同時に，ビブリオメトリクスは領域間や評価者間で，各評点の基準を明示的に共有することを支援することができる。ビブリオメトリクスにより，評価者は自分が評価を行っている少数のサンプル内で比較するだけではなく，「世界全体の論文の中で当該論文の引用回数が上位どのくらいに位置しているか」という分野横断的な指標を参照することができ，異なる分野の評価者の間でも大きな評価基準の相違が生じないように支援することを可能とする。

だが，繰り返し述べたようにビブリオメトリクス分析は論文や引用に関した一つの参照情報でしかなく，これがピアレビューに取って代わるものでは決してない。他方で，ピアレビューも冒頭に述べたように決して完全なものではない。そのため，ビブリオメトリクス以外にも，受賞数，特許数，招待講演数，競争的研究費の獲得額など様々な情報のもとでピアレビューは実施されるべきであり，これらの指標がピアレビューも含めてほぼ同様の結果を示す場合には，その評価結果の信頼性が高いことを期待できるのである（Martin and Irvine 1983, Martin 1996）。

他方，そもそも今回の事例である NIAD の評価は大学等の組織の改善の促進，および社会へのアカウンタビリティを目的とするものである。前者については，そもそも個々の教員レベルではなく，組織全体の特徴を示すことにも，ビブリオメトリクスがいかに用いられるかを検討する必要がある。例えば組織の研究活動の分野・研究テーマの広がりの特徴やその競争力を示したり，企業あるいは他国研究機関との共同研究関係といった，研究を実施してい

る構造の特徴を示すことで,組織の研究活動の改善や戦略形成への支援を行うべきである(林2001)。また，もう一つの目的であるアカウンタビリティに関しては，ピアレビューもビブリオメトリクス分析も評点を公表するだけでは不十分である。本来，公的資金を用いた活動をピアレビューにより評価するということは，一般社会から科学的・専門的事項の質の判断や保証を行う権限をその分野の専門家集団に委譲されたものと見ることができる。しかし，ピアレビューが評点や可否などの「専門家によるお墨付き」といったパターナリスティックな記号的情報しか提供しない場合には，逆に研究活動を行っている側と社会との間の情報量は少なくなり，アカウンタビリティの点ではマイナスに作用する可能性すらある。NIAD の評価では優れた研究内容についての記述を行うことでアカウンタビリティの確保への努力を行っている。ビブリオメトリクスにおいても，順位付けだけでなく，各研究がいかに優れているかを引用関係などから示すことが求められる。特に自己評価の中で，各組織が自己の研究活動の長所を明示的に示すことを可能とする方法として，このような定量的分析がいかに用いれるかを検討することが今後必要となる。

謝　辞

本分析を行うにあたり，本機構で理学系研究評価の運営を担当された川口昭彦評価研究部長，ならびに，評価事業部第３課の方々から多くの支援とコメントをいただいた。また，荒船次郎副機構長，芳鐘冬樹助手からも有益なコメントをいただいた。ここに記して感謝申し上げたい。

【参考文献】

Amin, M. and M. Mabe （2000）, “Impact factors: use and abuse”, *Perspective in Publishing*, No.1

Anderson, R.C., F. Narin and P. McAlister （1978）, “Publication ratings versus peer ratings of universities”, *Journal of the American Society for Information Science*, **29**, 91-103

Braun, T. （1999）, “Bibliometric indicators for the evaluation of universities – intelligence from the quantitation of the scientific literature” Scientometrics, 45, 425-432

Butler, L. （2003）, “Explaining Australia's increased share of ISI publications--the effects of a funding formula based on publication counts”, Research Policy, 32, 143-155 （2003）

Carter, A.M. （1966）, An assessment of Quality in Graduate Education, American Council on Education

Chubin, D.E. and E.J. Hackett （1990）, Peerless Science, SUNY Press

大学評価・学位授与機構（2002）『H12年度着手　分野別研究評価報告書集』http://www.niad.ac.jp/

Gibbons, M. and L. Georghiou （1987）, Evaluation of Research – a selection of current practices, OECD

Gibbons, M et al. （1994）, The New Production of Knowledge, SAGE Publications （小林信一・監訳（1997)『現代社会と知の創造』丸善）

林隆之（2001)「大学の研究活動の定量的プロフィールの形成」『研究・技術計画学会 第16回年次学術大会講演要旨集』379-382

Irvin, J.（1989),”Evaluation of scientific institutions”, The Evaluation of Scientific Research, John Wiley &Sons, 141-168

Kostoff, R.N. （1994） “Federal Research Impact Assessment: State-of-the-art”, Jouurnal of the American Society for Information Science, 45, 428-440

Kostoff, R.N. （1997）, Research Program Peer Review: Principles, Practices, Protocols

Martin, B.R. （1996）, “The use of multiple indicators in the assessment of basic research”, Scientometrics, 36, 343-362

Martin, B.R. and J. Irvine （1983）, “Assessing basic research”, Research Policy 12, 61-90

Merton, R. （1973）, The Sociology of Science - Theoretical and Empirical Investigations, The University of Chicago Press,.

Narin, F. et al. （1976）, Evaluating Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity, CHI

Office of Technology Assessment （1986）, Research Funding as an Investment: Can We Measure Returns?

Oppenheim, C. （1997）, “The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology”, Journal of Documentation, 53, 477-487

Oppenheim, C. （1995）, “The correlation between citation counts and the 1992 research assessment exercise ratings for British-library and information-science university departments”, Journal of Documentation, 51, 18-27

Polanyi, M. （1962）, “The republic of science –Its political and economic theory”, Minerva, 1, 54-73. 「科学の共和国」佐野安仁・澤田允夫・吉田謙二監訳（1985）『値と存在』晃洋書房

Rinia, E.J., Th.N. van Leeuwen, H.G. van Vuren and A.F.J. van Raan （1998）, “Comparative analysis of a set of bibliometric indicators and central peer review criteria”, Research Policy, 27, 95-107

Rinia, E.J., Th.N. van Leeuwen, H.G. van Vuren and A.F.J. van Raan （2003）, “Influence of interdisciplinarity on peer-review and bibliometric evaluation in physics research”, Research Policy, 30, 357-361

Roose, K.D. and C. Andersen （1970）, A Rating of Graduate Programs, American Council on Education

Shubert, A. and T. Braun （1996）, “Cross-field Normalization of Scientomerics Indicators” Scientometrics 36, 311-324

Swinbanks, D., R. Nathan and R. Triendl （1997）, “Western research assessment meets Asian cultures” Nature, 389, 113-117

Van Raan, A.F.J.（1996）, “Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises”, Scientometrics, 36, 397-420

Vinkler, P. （1987）, “An attempt of surveying and classifying bibliometric indicators for scientometric purposes”, Scientometrics, 13, 239-259

Zhu, J., A.J. Meadows and G. Mason （1991）, “Citations and departmental research ratings”, Scientometrics, 21, 171-179

[ABSTRACT]

Is Bibliometrics Useful to Support the Peer-review?
A Case Study of NIAD's Research Evaluation in Science

HAYASHI Takayuki*

In Japan, the lack of the quantitative information for peer-review in research evaluation has been criticized. Among them, bibliometrics indicators are the typical information to analyze the productivity and the impact of research, but it is unclear whether the western database can be used for research evaluation in Japan. In this study, the usefulness of bibliometrics in research evaluation in Japan is examined by the case-study of NIAD's research evaluation in science (mathematics, physics, chemistry, biology, geology and astronomy). The result shows that only half of the publications can be under analysis by the Science Citation Index in mathematics and geology, while 90% in physics, chemistry and biology. The positive correlation between peer-review and the citation indicator were found in all disciplines except mathematics. The comparison of the results of rating between peer-review and citation indicators showed that the two level gaps occur by only few percentages, so these gaps can ask reviewers to check their result again. And the normalized indicator of citation can be one of the information for the panels to adjust their criteria of each rank over reviewers and disciplines.

* Research Fellow, Faculty of University Evaluation and Researech, National Institution for Academic Degrees and University Evaluation