第 1 回 SPARC Japan セミナー2012

「学術評価を考える」

ビブリオメトリックスを活用した研究評価の現状と展望

孫 媛
（国立情報学研究所情報社会相関研究系准教授）

講演要旨

1980 年代以降、研究評価の重要性が増し、客観的評価のための指標への要望が世界的に高まっている。そうした要望に応えるものとして近年注目されているのが、ビブリオメトリックスで提案された諸指標である。ビブリオメトリックスは、図書館学・情報科学の一下位分野として始まったが、過去 40 年の間に徐々に、科学政策・研究マネジメントへと応用領域を広げてきた。そうした研究では、論文の書誌情報をもとに考案された一連の指標が活用されている。現在、ビブリオメトリックスの手法を研究評価に活用する方法を研究するためのセンターや評価室が、世界の多くの国々において創設されている。本講では、研究評価のツールとしてのビブリオメトリックスと研究評価の現状について概説する。ビブリオメトリックスの手法は有益なものではあるが、それを用いるだけですべての問題が解決するほど、研究評価の問題は単純なものではない。研究評価の今後の展望についても改めて考えてみたい。

孫 媛
1992年東京大学大学院教育学研究科教育情報科学専攻博士課程修了。学術情報センター助手，同助教授を経て，現在，国立情報学研究所，総合研究大学院大学准教授。専門はビブリオメトリックス（書誌計量学）・教育心理測定論。日本や世界の学術研究システムの実態と変化の様相を，ビブリオメトリックスの手法にもとづいて研究している。現在，関心を向けているテーマは，産官学間および国際的な連携，研究コミュニケーションネットワークの形成過程および大学研究力のベンチマーキングである。

研究評価とは

研究評価とは、研究活動に関する何らかの意思決定を行うために、評価対象の価値を判断する作業を指します。評価対象は大きく分けて、国、大学・研究機関、そして最近は研究者個人の評価も非常に多くなっています。また、プロジェクトや研究に関する政策、論文や雑誌などがあります。もともと研究評価は、投稿された論文を査読してジャーナルに掲載するか決め、研究者コミュニティ自らが公開されるべき研究成果を選別するところから始まりました。それにより、学問分野の質を高く維持する制度として広く定着してきました。つまり、研究評価の当初の目的は、優れた研究の探索と研究の学問的質の維持だったのです。それが 80 年代からは研究プロジェクトへの資金配分に用いられるようになり、だんだん科学政策と連動するようになりました。そして 90 年代にはその動きがさらに加速し、今は研究評価というと、大学ランキング、研究費配分、個人評価などが連想されるのではないかと

思います。

研究評価手法の変化

そうした社会情勢を背景に、研究評価手法も変わってきました。1970年代までは、研究対象を専門分野とする専門家が評価する同業者評価（ピアレビュー）という方法が主に用いられており、論文の査読や研究者への学術賞の授与、資金配分、プロジェクトの選定、人事などいろいろなレベルで行われていました。しかし1980年代以降は、客観的評価指標への要望が高まっていきました。

その背景には幾つかの要因が考えられます。まず、研究者数と研究活動全体が非常に拡大する一方、経済不況を背景に、国の研究予算の伸びは鈍化あるいは減少しました。それにより、すべての研究者や研究プロジェクトに資金を提供することができなくなり、評価することで選択的に資金配分する必要性が強まってきました。それに伴い、プロジェクト間や分野間の比較が必要になってきたため、専門家内でのピアレビューには限界が出てきたのです。また、研究活動に多大な公的資金が投入されているため、納税者への説明責任（accountability）が問われるようになりました。つまり、研究活動への公的資金投入は正当なのか、およびその効果は投入資金に見合っているかということをパフォーマンス指標で、専門家のみならず一般市民にも明確な形で分かりやすく示す必要があります。さらに、ピアレビューには非常に大きな問題点が指摘されています。それは、評価者の主観性や能力による意識的・無意識的なバイアスが生じる恐れがあることや非常に時間がかかるという点です。科研費の審査官をするときも結構苦労すると思いますが、専門内のものであっても大変時間がかかります。また、実際に評価をするとコストも非常にかかります。

ビブリオメトリックスとその略史

そうした状況の中で、ビブリオメトリックス指標による研究評価が増えました。ビブリオメトリックスは、もともと図書館情報学の一つの研究ツールでしたが、それが評価ツールとして定着していきました。

ビブリオメトリックスという言葉は、図書をはじめとするコミュニケーションメディアに対する数学と統計的方法の適用を定義として、1969年にAlan Pritchardがはじめて用いたものです。文献データの統計分析手法は、科学コミュニケーションの解明・測定にも用いられることから、サイエントメトリックスの同義語として使われることが多いです。分析対象は主に論文ですが、最近は特許やウェブデータなど、多岐にわたっています。

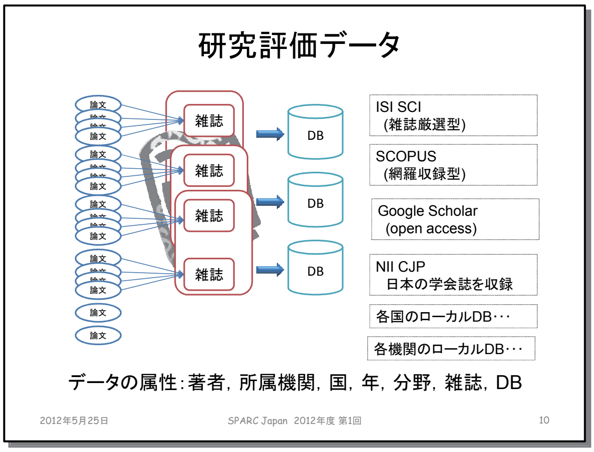
ビブリオメトリックスにはおおむね50年の歴史があります。ビブリオメトリックスという言葉は1969年に生まれたのですが、研究はそれ以前からされており、1917年のColeの論文が、評価を意識した初めてのビブリオメトリックス分析と言われています。歴史の中で最も重要な出来事は、Eugene Garfieldが60年代にSCI（Science Citation Index）を創設したことです。その結果、引用統計の調査が可能となり、ビブリオメトリックスの応用研究の可能性が飛躍的に拡大しました。1963年にはPriceが著書『Little Science, Big Science』の中で、科学コミュニケーションの分析を通して、ビブリオメトリックスの系統的手法を提示し、論文数・引用数を用いる現在の研究評価手法の基礎を築きました。1970年代には、Henry Smallが共引用分析という手法、またFrancis Narinがビブリオメトリックス指標を初めて研究評価・科学政策に適用し、ピアレビューとの相関を示しました。

その後、論文データベースを用いて論文や引用を分析することにより、科学や科学コミュニケーション、知識生産の特徴を明らかにできる可能性が非常に大きいという認識が研究者の間で次第に定着し、研究評価にビブリオメトリックスを活用する研究グループが次々と出来上がりました。代表的な研究センターとして、サセックス大学のSPRU（イギリス）、ライデン大学の CWTS（オランダ）、CPI（オーストラリア）、CINDOC（スペイン）などが挙げられます。

その後、アメリカの Office of Technology Assessment が 1986 年に出したテクニカル・メモランダムの中で、ビブリオメトリックス指標をピアレビューと併用することの重要性と可能性を提唱しました。これにより、主観的指標が研究評価を支配していた時代が終わったと言えます。

ビブリオメトリックス指標の研究評価への利用

最近の評価データは論文に限らず、プロシーディングスも少し入るのですが、主に論文を雑誌に投稿して、その雑誌が Web of Science や Scopus などのデータベースに収録されます（図 1）。Web of Science は雑誌厳選型、Scopus は網羅収録型と言われており、類似点と相違点があります。この両者は有料でだれでも使えるわけではないのに対し、Google Scholar は無料なため、最近非常に利用が増えています。また、NII の CJP や各国・各機関のローカルデータベースも評価データになります。


研究評価データ

ISI SCI
（雑誌厳選型）

SCOPUS
（網羅収録型）

Google Scholar
(open access)

NII CJP
日本の学会誌を収録

各国のローカルDB･･･

各機関のローカルDB･･･

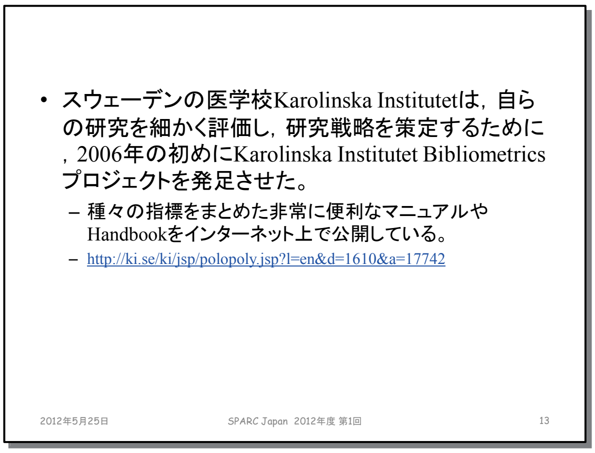
データの属性：著者，所属機関，国，年，分野，雑誌，DB

2012年5月25日　SPARC Japan 2012年度 第1回　10

（図 1）

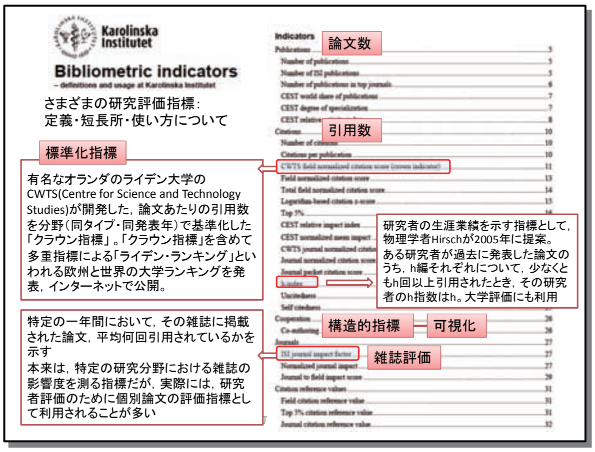
ビブリオメトリックスの最も基本的な指標は論文数と被引用数で、それぞれ研究の生産性や活発さを表します。グレードの高い雑誌への掲載は一定の質を保証すると考えられ、生産数だけではなく、ある程度質も表せます。また、この二つの基本指標を基に、多様な指標が構成されています。雑誌評価指標については JIF（Journal Impact Factors）や Scopus の SNIP などがあり、そのほかにも h-index などいろいろな指標があります。

それぞれの指標の特徴や使い方、注意点についてまとめられたマニュアルが、インターネット上で公開されています（図 2）。そのサイトには、論文数や引用数、CWTS 評価の標準化指標であるクラウン・インディケーター、h-index、雑誌評価指標、インパクトファクター、構造的指標などが載っています。構造的指標とは、論文の共著関係や共引用関係を可視化することによって、分野間の関係を見るものです。それ以外にも、国と国の研究ネットワークなどの分析指標がたくさんあります（図 3）。


- スウェーデンの医学校Karolinska Institutetは，自らの研究を細かく評価し，研究戦略を策定するために，2006年の初めにKarolinska Institutet Bibliometricsプロジェクトを発足させた。
  - 種々の指標をまとめた非常に便利なマニュアルやHandbookをインターネット上で公開している。
  - http://ki.se/ki/jsp/polopoly.jsp?l=en&d=1610&a=17742

2012年5月25日　SPARC Japan 2012年度 第1回　13

（図 2）



さまざまの研究評価指標：
定義・短長所・使い方について

標準化指標

有名なオランダのライデン大学のCWTS(Centre for Science and Technology Studies)が開発した，論文あたりの引用数を分野（同タイプ・同発表年）で基準化した「クラウン指標」。「クラウン指標」を含めて多重指標による「ライデン・ランキング」といわれる欧州と世界の大学ランキングを発表，インターネットで公開。

特定の一年間において，その雑誌に掲載された論文，平均何回引用されているかを示す
本来は，特定の研究分野における雑誌の影響度を測る指標だが，実際には，研究者評価のために個別論文の評価指標として利用されることが多い

研究者の生涯業績を示す指標として，物理学者Hirschが2005年に提案。
ある研究者が過去に発表した論文のうち，h編それぞれについて，少なくともh回以上引用されたとき，その研究者のh指数はh。大学評価にも利用

（図 3）

基本指標とその他の主な指標

論文評価の基本指標としては、論文数が非常に単純

で分かりやすいためよく使われています。特に研究機関や大学を評価する場合、論文数だけではなく、研究者一人当たりの論文数、高引用論文数（良い雑誌に掲載された論文数）、論文数の加重（重み付けのあり・なし）などの指標も利用されています。もう一つの基本指標は被引用数です。被引用数とは、論文が発表された時点からある時点までほかの論文から引用された回数です。これは論文の質に関する最大の客観的指標だと考えられています。このほかにも最近は代替指標（alternative indicators）というものがあります。国や機関評価の場合、研究者一人当たりの被引用数や1論文当たりの被引用数という指標もあります。研究者一人当たりの被引用数は簡単そうに見えますが、国を比較する際には研究者数のカウントの方法論から始まり複雑なため、実は非常に難しい指標なのです。

この二つの指標を基に開発されたそれ以外の指標として、インパクトファクターというものがあります。これはトムソン・ロイターが作成した引用索引データベースの収録雑誌について、掲載された論文1編当たりの被引用数の平均を算出したもので、もともとはSCIやCurrent Contentsへの収録雑誌を選定するための定量的基準として考えられた指標です。研究者個人や研究機関の評価指標に転用されることがよくありますが、これはそもそもそのために作られた指標ではないため、利用するときには注意しなければいけません。

それから、最近流行しているh-indexという指標があります。日本ではまだそうでもないかもしれませんが、海外では「私のh-indexはいくつか」という質問をされることが多く、図書館員が困っているという話を聞いたことがあります。h-indexは研究者の生涯業績を示す指標として、物理学者のHirschが2005年に提案しました。ある研究者が発表した論文のうち、少なくとも h 回以上引用された論文が h 編あるとき、その研究者のh-indexはhであるということです。例えば被引用数が10回以上の論文を10編持っている研究者の場合、その人のh-indexは10です。被引用数10以上の論文が9編しかない場合、h-indexは9になります。つまり、高い指標を得るためには、論文数も引用数も必要で、量と質の両立が求められます。この指標は発表当初から多くの研究者、政策立案者、メディアに非常に注目され、隔年開催のビブリオメトリックスの国際会議でもh-indexのセッションが増えており、h-indexの議論をする場が非常に多くなってきています。これを特に有名にさせたのはWeb of ScienceとScopusで、何年か前からh-indexが実際に出せるようになって以来、非常によく使われています。最近は過剰に使われているので、「何でもかんでもh-indexはおかしいだろう」という指摘もあります。

イギリスのRAE（Research Assessment Exercise）は、かつてはピアレビューしかありませんでしたが、だんだんメトリックスのみになってきました。近年は、研究評価の多様性が増す傾向にあり、メトリックスを主としてピアレビューを従とするもの、あるいはピアレビューの参考資料としてメトリックスに提供してもらうなどさまざまです。

ビブリオメトリックス指標を用いることの意義と留意事項

これらの指標を用いることにより、客観的指標への社会的な要請に応えることができます。例えば、ピアレビューの問題点を補完すること、説明責任に応えること、そして分野間・機関間・国間のグローバルな比較が可能になります。

しかし、留意点を自覚せずに使うと、誤った結論を導く危険もあります。留意事項①は、技術的な問題です。つまり、論文数・被引用数を正確に求めること自体、膨大な手間を要する作業だということです。また、データベースには著者の記述どおりの情報が収録されていることが多いのですが、例えば引用文献であれば表記法、雑誌名の省略法の違い、書誌情報記載の誤りなどにより、同じ論文であっても別のものとしてカウントされていることもあります。さらに、名前がイニシャルで表記されているだけでは、研究者の同定が難

しいこともあります。Web of Science と Scopus では書誌情報が基本的に英語で書かれているため、機関名の表記揺れや組織改組などによる機関名の変化が必ずしもタイムリーに反映されないという問題もあります。これらは「名寄せ」の問題と呼ばれています。最近は、「名寄せ」の技術やリサーチ ID などの対策が考えられています。

留意事項②は、論文数や被引用数にはいろいろな要因が影響を与えていることです。まず、データベースにより、雑誌の採録範囲が違います。例えば Web of Science や Scopus にはそれぞれに特徴があり、同じ論文の被引用数をそれぞれのデータベースで引くと、違う数字が出てきます。これはそもそも雑誌の採録範囲が違うので、当然のことです。また、引用習慣や論文の寿命などは、研究分野間で差異があります。たとえば、医学・ライフサイエンス分野では、発表から年数があまり経過していない文献を多数引用するのが普通であるため、一論文あたりの平均引用回数が多くなる一方で論文寿命は比較的に短いのに対し、数学の分野では、引用数はそれほど多くないけれども論文寿命は長いという特徴があります。したがって、統計を取る期間によっても、被引用数は影響を受けます。また、レビュー論文であるのかオリジナル論文か、英語の論文かどうかによっても引用されやすさが違います。

留意事項③として、共著論文の論文数・被引用数の統計の取り方についても、集計法の問題があります。その背景には、最近は単著がどんどん減り、共著・国際共著が非常に増えてきている、あるいは自己引用や否定的な引用があります。否定的な引用とは、ある論文の内容を非難しているのにカウントされるというものです。

留意事項④は、指標の特徴と固有の限界です。まず、評価対象に適合するかが非常に重要です。測ろうとしているものに対し、違うメジャーではなく、きちんと測れるメジャーを持っていかなければ問題になるので、そこはきちんと考えなければいけません。また、分野間の違いや引用統計の問題も指摘されています。例えばほとんどの論文は引用が少ないけれども、たまに非常に引用の高い論文があります。統計学では分布によって使える指標が違うとよく言われますが、非常にゆがんだ分布に対しては平均値を使うと不都合です。どうしても高い値に引っ張られてしまって平均値が上がり、実際の分布を反映していないからです。また、引用年数の問題などもあります。

留意事項⑤は、データベースへの依存性です。Web of Science や Scopus は有料ですが、Google Scholar は無料で、いつでも誰でも使えます。しかも、Web of Science や Scopus と同様、引用のトラッキングができるといった機能も付いています。さらに、Web of Science や Scopus は英文の文献だけですが、Google Scholar は論文、会議録、図書、各国の英語ローカルの論文も全部拾えます。このように、Google Scholar は無料でカバー範囲が広いため、使う人が非常に増えています。ビブリオメトリックスの学会でも必ず Google Scholar と Web of Science、Scopus を比較する研究があります。しかし、Google Scholar では、範囲と基準がどうなっているかは不明確です。適当に拾ってきている、あるいは二重、三重にカウントしているかもしれず、質の保証がないため、注意して使用しなければなりません。また、評価指標は各データベース内で計算されているため、異なるデータベースに基づいた指標間の比較ができません。

留意事項⑥は、分野分類法の問題です。Scopus と Web of Science では分野分類法が違います。また、論文ごとではなく、掲載雑誌ごとに分類が行われているので、分野ごとに評価をする際に問題が生じる可能性があります。最近はデータ統合の動きがあるのですが、そこでも分野分類の違いによる問題が生じています。

留意事項⑦は、概念・指標の再解釈の問題です。ビブリオメトリックスは研究評価・ベンチマーキングのツールとして、従来のビブリオメトリックスに新しい見方をもたらしているため、概念・指標に対して再解釈を行う必要がある場合があります。その再解釈自体

に必ずしも筋が通っていないこともあれば、解釈・利用することにより評価対象そのものを変えてしまうこともあります。例えば「引用」という概念に対する再解釈と効果について、従来のビブリオメトリックスでは、「引用＝情報の利用」です。つまり、以前は「引用なし＝情報が利用されていない」「引用回数が多い＝よく認識されている」「自己引用＝研究の一環」と考えられていました。現在の研究評価の文脈では、「引用＝質の評価」と解釈されています。そうすると、「引用なし＝質が低い」「引用回数が多い＝研究の質が高い」「自己引用＝Impact の歪曲」となります。すなわち、自己引用をしてはいけない、あるいは評価をするものから除くべきだと考えられています。これを再解釈することにより、実際研究者の引用行為をゆがめていると言えます。

研究評価のこれから－評価データの問題

今後の課題の一つとして、評価データの問題が挙げられます。ビブリオメトリックスの指標を持ち込む本来の目的は客観的に見ることなので、データは最も重要な問題の一つになります。現在、いくつかのデータベースがあり、収録方針や特徴はそれぞれ違いますが、共通の問題を抱えています。まず、雑誌の収録範囲が異なっています。特に人文社会分野の収録は少ないです。そして、文献の種別は、会議録や図書などいろいろありますが、ほとんどが雑誌です。また、国ごとでも収録雑誌数が違います。例えば Web of Science には日本の雑誌は約 200 誌、Scopus には 600～700 誌あると思います。それから、英語の雑誌が圧倒的に多いという問題や、分野分類体系が違うという問題があります。さらに、バイアスをなくし、公平な評価や目的に合った評価をどのように行えばいいのかを考えなければいけません。

評価は大きく分けて、各国内でのローカル評価とグローバルな評価があると思います。最近は、ファンディングと連動させたい、S&T Innovation system の中に産学連携・研究者情報を入れたい、あるいはその情報を利用したい、自分の国や機関の研究に関する現状を把握したいなどのローカル評価のニーズが非常に高まっており、いろいろな動きがあります。それに対して、自国の雑誌が網羅されておらず、非常に大きな問題になっています。人文社会科学・応用分野のデータが不十分だという現状もあります。また、グローバル評価では、国間の研究プロフィール・ネットワークの分析、世界大学ランキング・ベンチマーキングといった動きがありますが、使っているデータの偏りが問題になっています。

そこで、最近はローカル評価用データの整備が非常に進んでいます。既存のデータベースを拡張して、統合的なデータベースの作成が試みられています。例えばオランダの CWTS やベルギーの SOOI、Web of Science、Scopus などは、既存のデータベースに自国のデータを入れ、全体として評価しています。また、最初からデータベースを作ってしまうという動きもあります。例えばオランダとベルギーのフランダース地方では、VABB-SHW という人文社会科学のデータベースを作っています。しかし問題もたくさんあります。まず、基準の違いや研究者別の割合の違いなどがあり、データを混ぜても指標が統一されていません。また、新しくデータベースを作っても、実際に指標が使えるようになるまで非常に時間がかかります。

さらに、自国の引用索引データベースを作るという動きもあります。日本では 1995 年に NII が作りました。中国ではそれよりも少し前に、中国版の引用索引データベースを作りました。NII が作っているものには人文系は含まれていないのですが、中国の場合は人文系も含んでかなり力を入れてやっており、Web of Science のプラットフォームの中にも載せています。韓国、台湾、タイでも 2000 年以降、そうした動きがあります。マレーシアでは、国が多くの予算を付けて、自国の引用索引データベースのセンターを設立しました。さらに、最近は東南アジア諸国連合（ASEAN）引用データベースセンターを立ち上げてそれぞれの国のデータベースを一緒にグルーピングして利用するこ

とで、国のプレゼンスを高めようという動きもあります。ただ、実際に尺度の違う指標を比較する場合は、尺度の等化を行うことが必要です。

また、グローバル評価用データの問題もあります。大学ランキングの結果などは、データにバイアスがあり、DB 間の結果を比較することが難しいという問題もあるので、使用した結果を解釈するときに、十分そのことを認識する必要があると思います。

新しい指標の研究開発

ランキング結果を出そうとするとき、いろいろな側面があるにもかかわらず、指標を重み付けしたり、加工をしたりして、一つの尺度にまとめようとします。それではやはり無理があるので、最近は多次元評価という考え方が出てきました。また、ランキングではなく、お互いに比較対象と比較しながらベンチマーキングするという動きや、代替指標の開発もあります。

今日は、雑誌のオープンアクセスや電子ジャーナル、機関リポジトリ、研究者自身による論文のウェブ公開といった学術情報をめぐる環境が激変しているため、今後も評価対象や指標は変わっていく可能性があります。新しい環境に対応する評価指標の研究開発も行われています。研究評価にはいろいろな側面があるので、まず何をどう評価したいかをよく考える必要があります。そして、既存データにとらわれずに、いろいろなデータを使うべきだと思います。研究評価に関する技術については進展が期待できますが、今後の研究形態の大きな変化に伴い、評価指標も変わらざるを得ないと思います。

研究評価は科学政策に資するのか

評価は非常にコストがかかります。お金の面だけでなく、研究者にも負担がかかり、いろいろなビヘイビアまで変わることもあるので、それに見合う成果が得られているかを考える必要があります。しかし、評価自体がいいか悪いかは別として、社会からの要請がある以上、研究評価をしなければなりません。ビブリオメトリックス研究者としては、新しい要請や環境に合わせて、新しい技術や指標の開発・改良をしていく努力をするだけです。しかし、追いつけない現状もあるため、使う側としては指標の限界や注意点を十分に認識する必要があります。

研究評価に今後求められる方向性

研究評価について今後求められる方向性の一つは、評価する側だけが情報・利益を受けるのではなく、評価される側にとっても意味のある評価という方向性だろうと、私は思っています。私自身のもともとの専門は心理測定・評価です。いまでも、教育・心理テスト理論の研究も行なっていて、最近の主な研究テーマは、ネット社会における認知診断テストです。最後に、教育評価研究との比喩で研究評価をみてみたいと思います。

教育評価は中国の科挙制度まで遡ることができますが、古くは「文章を書かせる」「口頭試問」のように、専門性を持った試験官が主観的に判断するものでした。その後、20 世紀の初頭以降、そうした評価法の主観性・バイアス等が指摘されて、多肢選択式などの客観テストが主流になり、テスト理論という専門分野が生まれました。客観テストデータの解析法について盛んに研究されるようになったのが 20 世紀半ば以降のことです。テストをする側からすると客観性はとてもありがたいのですが、ともすれば評価のための評価になり、本来の目的であるはずの「教育」の側面が軽視されているという反省も出てきたのです。近年は評価と指導は一体でなければならないということが言われるようになっています。私が研究している認知診断テストも、学習者を客観的に序列化することが目的ではなく、学習者のつまずきの原因を探り出し、学習者自身の学習を支援する形で情報を用いることを主な目的としているのです。

研究評価も同様な道を歩んでいるような気がします。ピアレビュー、メトリックス、多様な方法、さらには研究・機関の意思決定の支援の提供という方向になっ

ています。今後どうなるのか、もちろん具体的な指標などについても、まだまだ改善・進歩の余地はありますが、目指すべき一つの方向性は、評価される側が自分たちのために活用しうる評価ということになるだろうと思うのです。