(DRAFT—please do not cite or redistribute)

The Ethics of Evaluative Bibliometrics

Jonathan Furner

Introduction

The purpose of this chapter is to present a theoretical framework for the study of the ethical aspects of evaluative bibliometrics. The practice of evaluative bibliometrics involves the use of quantitative methods to analyze the decisions made by authors and readers of documents, and the use of the results of that analysis to inform decision making in the processes by which authors are rewarded for their work. Expressions of partial or complete theoretical frameworks for the study of bibliometric practice abound in the literature, but few provide foundations appropriate for study of its ethical dimension. This chapter is intended to fill that gap.

In the first two sections to follow this introduction, evaluative bibliometrics is situated in the context of its overlapping parent fields of

bibliometrics and research evaluation, and a version is presented of a theoretical framework for bibliometrics that, as is typical, omits ethical categories. A justification is then provided for the decision to focus on the ethical dimension. The existence of such a dimension is demonstrated, its content and scope defined, and its significance evaluated. The next section provides the chapter's primary contribution: a framework for the study of bibliometric ethics. The values and principles of participants in evaluative studies are reviewed, and the lack of community-wide consensus on principles of distributive justice is highlighted as a core concern. The chapter concludes with a remark on the potential for applying a framework of the kind developed here to the study of other types of uses of bibliometric techniques.

Bibliometrics and Evaluation

Bibliometrics has been defined as "the study of the quantitative aspects of the production, dissemination, and use of recorded information" (Tague-Sutcliffe, 1992, p. 1; quoted in Bar-Ilan, 2010, p. 2755). More informally, we might say that bibliometrics is about what *people* (authors, readers, etc.) do with *documents* (books, journal articles, web pages, tweets, etc.), for what reasons, and with what effects. It involves the observation, classification, and counting of document-related actions (writing, submitting, reviewing, editing, publishing, viewing,

buying, reading, citing, etc.), and the ranking and mapping of classes of such actions, in order to produce representations of patterns and trends in document-related behavior. These representations, in the form of descriptions and indicators of various numerical and graphical kinds, can in turn be used to (a) *reward* people for their past activity as authors or readers; (b) *recommend* particular documents, or classes of document, for future use; or (c) simply improve our *understanding* of the processes underlying the structures and dynamics of networks of documents and related entities.

An assumption at the core of this conception of the nature and scope of the field of bibliometrics is that any document-related action of the kinds listed above is the outcome of a decision to select, at time *t*, one particular document (or class of document) rather than any other as the object of the action. In other words, the action is treated as an expression of a preference ordering over the universal set of documents. Analysis of multiple such preference orderings allows us to produce (a) composite *rankings* of documents (or of classes of documents), which may then be used as the basis for rewarding the authors of highly ranked documents, and/or recommending highly ranked documents to information seekers, and (b) *maps* or graphs showing the relationships among documents (or among classes of documents), which may then be used as the basis for recommending strongly

related documents to information seekers, and/or representing or describing the structure of document networks.

Evaluative bibliometrics (see, e.g., Narin, 1976) is the branch of the field that focuses on (a) the specification of techniques for the production of rankings, and (b) the use of such rankings as the bases for distributing resources or credit among the individuals responsible for ranked documents, or among the institutions with which authors are affiliated. University administrators use the techniques of evaluative bibliometrics in faculty tenure cases, in the course of identifying authors deemed most worthy of promotion; government agencies use evaluative bibliometrics in the allocation of research funding, in the course of identifying departments, programs, and projects deemed most worthy of support (see Lane and colleagues, chapter 21, this volume); librarians use evaluative bibliometrics in collection development, in the course of identifying journals deemed most worthy of purchase or licensing for access by library users (see Haustein, chapter 17, this volume).

Viewed as a set of techniques, evaluative bibliometrics is just one of several options available to would-be evaluators of research and/or researchers. The distinct but overlapping subfield of *research evaluation* (see, e.g., Whitley & Gläser, 2007) is dedicated to the study and application of such sets of procedures for the systematic determination of the value[1] of research projects and programs,

of their outputs and outcomes, and of those who lead and participate in them. Research evaluation is itself a branch of the field of *evaluation* (see, e.g., Scriven, 1991), whose practitioners inquire into the general process of determining the value of agents, objects, events, etc. (i.e., "evaluands"), of any given kinds. An important result of work in the latter field is an outline of a general procedure for evaluation that involves the following tasks:

- Specification of the *variables* (aka properties, states, conditions, qualities, attributes, criteria, dimensions) whose values[2] are to be used to characterize evaluands
- Specification of the methods to be used of *operationalizing* the chosen variables so that measurements may be taken easily and reliably
- Specification of the methods to be used of *normalizing* values of the chosen variables so that measurements taken under different conditions (e.g., over different time periods) are comparable
- Optionally, specification of the methods to be used of *weighting* the chosen variables so that measurements may be combined in a single, overall metric

Justifications of particular choices of variables may make claims for the *intrinsic value* (sometimes known as merit) of selected variables, and/or for their instrumental or *extrinsic value* (sometimes known as worth or "goodness-for"). Justifications of the latter type may include additional specification of the

purposes, goals, or *functions* of evaluands for the members of one or more groups of *stakeholders*.

Evaluations are themselves undertaken for a variety of purposes. The instigators of evaluative studies may be primarily interested simply in knowing how the evaluands in a given population compare with one another. They may also wish to use the results of an evaluation as warrant or grounds for choosing among evaluands, or for allocating varying quantities of resources or rewards to different evaluands. Alternatively, they may wish to determine how the value of evaluands might be improved, or to encourage evaluands to consider the fact of evaluation (or the prospect of reward) as a motivation or incentive to achieve their goals more successfully. Finally, administrators may consider it their duty or responsibility to undertake an evaluation in order to meet professional standards of accountability.

Together with the effects typically intended by administrators—better decision making, fairer allocation of resources, improved performance and/or reputation—evaluative studies can also have *unintended side effects* of various kinds. Evaluands might see their involvement in the evaluation, or their expectation as to its outcome, as an incentive to change their behavior with results that run counter to those desired by administrators. Methods of evaluation may themselves be treated as evaluands, and their intrinsic and extrinsic value

determined in a process of *metaevaluation* in which undesired effects are set against desired ones.

Two types of analysis, distinguished by the variable on which subjects are assessed, are dominant in evaluative bibliometrics. *Publication analysis* is based on counts of the occasions on which the documents produced by each author (or by each organization, each subject area, each country, etc.) have been published; *usage analysis* is based on counts of the occasions on which the documents produced by each author (or by each organization, each subject area, each country, etc.) have been used.[3] Citation analysis is a specialized form of usage analysis in which it is assumed that counts of citations serve as reliable evidence of the amount of use to which citing authors have put cited documents. Usage analysis is itself sometimes conceived of as a form of *impact analysis*, on the assumption that counts of usage events (i.e., citations, links, loans, holdings, downloads, views, etc.) serve as reliable indicators of the amount of impact that documents have had on a given population of users. Similarly, publication analysis is sometimes conceived of as a form of *productivity analysis*, on the assumption that counts of publications serve as reliable indicators of the rate at which their authors are productive.

Our assessment of the validity of analysis of each of these kinds rests on our attitudes toward each of a chain of successively more basic premises: the

claim that values of the chosen variable of evaluation (rate of productiveness, amount of impact, etc.) are positively correlated with measurements of the level of *quality* (i.e., the goodness) of research, and thus that rankings derived from publication and/or citation counts can be used as surrogates for measures of quality; the belief that the quality of research is the most appropriate basis on which to assess the extent to which researchers are *deserving* of reward; and the belief that desert[4] is the most appropriate basis on which to distribute reward. In summary, arguments in justification of the validity of using bibliometric techniques in research evaluation need to demonstrate (a) that publications are evidence of productivity and citations are evidence of impact; (b) that productivity and impact are evidence of quality; (c) that quality is the appropriate basis for the assessment of desert; and (d) that desert is the appropriate basis for the distribution of reward.

A Conceptual Framework for Bibliometrics

As one would expect, the literature of bibliometrics is vibrant and multifaceted, replete with contributions to many different debates on methodological and other foundational issues, as well as with reports of the findings of studies in which bibliometric techniques have been applied (see, e.g., Bar-Ilan, 2008; Borgman & Furner, 2002).[5] A framework for classifying the most significant foundational

issues might include the following categories, among others. Taken in combination, contributions in these categories allow for detailed description and explanation of the nature and scope of the field, of subfields such as evaluative bibliometrics, of the distinctions between bibliometrics and related areas of inquiry, and of its disciplinary affiliations:

- *Purposes*: specification of the general kinds of questions, problems, and issues, and of the particular instances of those kinds, that bibliometricians seek to answer, resolve, or understand
- *Uses*: specification of the kinds of contexts and environments, and the kinds of ways in which the outcomes of bibliometric research may be applied
- *Ontology*: clarification of the commitments that bibliometricians have to the existence, in reality, of entities in various fundamental categories
- *Epistemology*: clarification of the processes by which bibliometricians believe it is possible to acquire knowledge of the subject matter of bibliometrics
- *Methodology*: generally, specification of the methods by which valid and reliable data may be collected, and of the methods by which relevant and appropriate analysis of data may be carried out
- *Metamethodology*: explanation and evaluation of the general approaches that may be taken, and the particular methods that may be used, to address the foundational issues listed above

- *Paradigms:* at the most general level, identification of the paradigms within which bibliometricians may (consciously or unconsciously) operate

Turning to focus on methodology in particular, we find contributions of the following kinds, most of which are generic to fields that involve the development and application of statistical techniques:

- Specification of the kinds of *phenomena* (objects, properties, actions, agents, etc.) about which data may be collected and analyzed for bibliometric purposes, and of the *levels* or units at which phenomena may usefully be aggregated and analyzed
- Specification of the kinds of *data* that may serve as evidence of the influences on and/or effects of human document-related activity
- Specification of the kinds of *observation* required to produce data that are valid and reliable indicators of the existence of structures and operation of processes
- Specification of the methods by which *descriptions* of sets of bibliometric data are produced in the form of summary statistics (aka metrics, indicators), ranked lists, and graphical visualizations
- Specification of the methods by which mathematical *functions* are generated as putative descriptions of the regularities found in distributions of the probabilities of occurrence of observed phenomena

- Specification of the methods by which we may calculate the *goodness of fit*, to the data collected, of the functions proposed
- Specification of the methods by which *models and theories* may be produced as explanations of regularities
- Specification of the methods by which we may *evaluate* the utility, coherence, and/or correspondence with reality, of the models and theories proposed as explanations of regularities
- Specification of the kinds of *technologies and tools* that may be used to support efficient and effective data collection and analysis

Lastly, consideration of aspects most germane to evaluative bibliometrics leads to the following list of the kinds of choices among available alternatives that must be made and justified by analysts working on any given evaluative study:

- Selection of the *unit type(s)* of evaluands to be studied: e.g., documents, authors, journals, departments, institutions, nations, fields
- Selection of a method of identifying the particular *population(s)* of evaluands to be studied: e.g., institutional membership, database coverage
- Selection of the *variable(s)* whose values are to be used to characterize evaluands: e.g., productivity, impact on science/scholarship, impact on society, research quality, equality, diversity

- Selection of a method of *operationalizing* the chosen variables so that measurements may be taken: e.g., counting publications, counting citations
- Selection of a method of *normalizing* values of the chosen variables so that measurements are comparable: e.g., by time period, by frequency of citable documents
- Selection of a method of *weighting* the chosen variables so that measurements may be combined in a single, overall metric
- Selection of a method of *ranking* normalized values of operationalized variables for the evaluands in the chosen population

Neither the general process, nor the specific outcome presented above, of constructing a theoretical framework for evaluative bibliometrics along these lines could be construed as a novel contribution. The level of detail is necessary, however, to demonstrate a significant omission: the *ethical* dimension, which (I claim) cuts across many of the categories listed. Treating evaluative bibliometrics as a discrete set of techniques for the evaluation of the agents and products of authorship, we may engage in a form of metaevaluation in which we determine the intrinsic and extrinsic value of the kinds of methodological choices made in each of the given categories. Extrinsic value is assessable relative to the goals of stakeholders, but how might we go about measuring intrinsic value (of choices

and/or the goals of choosers)? This is where a foray into the field of ethics is helpful.

Ethics, Values, and Principles

Ethics (see, e.g., Shafer-Landau, 2010) is the area of inquiry, normally treated as a branch of philosophy, in which answers are sought to questions like "What is the right thing to do?" and in which methods and results of thinking and reasoning about such questions are studied and evaluated. Well-established subfields of ethics include *normative ethics*, which is productive of specifications of criteria for distinguishing between right and wrong actions, and of theories that provide justifications for those specifications; *metaethics*, which is productive of methods of classifying ethical theories; and *applied ethics*, which is productive of demonstrations of the consequences of applying criteria of particular kinds as guides to action in situations of particular kinds.

Professional ethics is that subbranch of applied ethics concerned with the ethical aspects of work in the various professions.[6] A tool found to be useful by the leaders of many professional associations is the *code of ethics*, which can take any of a variety of forms (and a variety of titles) but the primary purpose of which is typically intended to be to ensure that members of the given profession have the opportunity (by studying the code) to develop an awareness and understanding of

the kinds of practices generally considered by their peers to be justifiable by ethical principles. Secondary purposes of codes of ethics include (a) the communication of the values of the profession to nonmembers, that is, to the consumers of the goods and/or clients of the services provided by members of the association, as well as to policymakers, journalists, and members of the public; and (b) the establishing of a means of holding members of the profession to account for actions perceived not to be justifiable by ethical principles.

The forms taken by codes of ethics do vary, but one structure commonly adopted involves a distinction being made between statements of the profession's values, and statements of principles. *Values* are those kinds of states, conditions, properties, etc.—variously attributable to agents, objects, events, or other phenomena, as individuals or in aggregations—that (it is claimed) are *good*. *Principles* are specifications of the kinds of conditions that must be satisfied, the kinds of states that must prevail, the kinds of properties that must be instantiated, for any given action to be deemed *right*.

When a person is said to "hold" a certain value, then the claim is that that person believes that a certain kind of state, property, etc., is good. Definitions of goodness proliferate, as do typologies of kinds of goodness, but one feature commonly (if only implicitly) attributed to goodness is its quantifiability, in one or both of two senses: all other things being equal, the more we have of a good

thing, the better; and again, all other things being equal, the more things we have that are good, the better.

Different ethical theories propose different kinds of justification for action-guiding principles, and different conceptions of the ways principles relate to values. According to theories in a family known as consequentialism, for example, principles are justified to the extent that the actions they recommend tend to produce effects characterized by greater quantities of values. The rightness of actions, in other words, is determined by the goodness of their *consequences*. Such theories suggest that, if we are interested in the possibility of a better world, it is rational for us to act in whatever way is productive of higher frequencies of occurrence of those states, properties, etc., that we identify as values. On a view of this kind, principles may be treated as specifications (ranging from the very general to the very specific) of the kinds of actions that (it is claimed) have higher probabilities than do alternatives of producing greater quantities of values.

Other theories propose justifications for principles that pay less attention to the goodness of the consequences of the actions recommended by those principles, and more to the goodness of the *reasons* that agents have for acting in those ways. On some views of this kind, values may be treated as virtues attributable to agents, and principles as specifications of the kinds of actions that tend to be characteristic of virtuous agents.

A Conceptual Framework for Bibliometric Ethics

However we decide to theorize the relationship between values and principles, it appears that one productive way we might structure an inquiry into the ethics of evaluative bibliometrics would be to focus on the following tasks:

1. Identification of relevant *subgroups*, each distinguished by their members' shared goals, of the population of agents responsible for actions taken in the course of bibliometric evaluations
2. Identification of the kinds of *actions* taken by the members of each subgroup in the course of bibliometric evaluations
3. Identification of the *values* held by the members of each subgroup
4. Identification of the *principles* for which the members of each subgroup advocate
5. Identification of *holes* in the ethical systems analyzed, where guiding principles would be useful and yet are absent
6. Identification of *violations*—that is, activities indicative of the values and/or principles of one subgroup lacking correspondence, or coming into conflict, with those of another

Subgroups

The three subgroups of agents at the heart of any bibliometric evaluation are as follows:

- *Analysts*: i.e., the bibliometricians responsible for collecting and analyzing data on the document-related activities of specific groups of subjects and reporting on their findings
- *Users*: i.e., the administrators and policymakers responsible for commissioning bibliometric studies, and for using the results of such studies to inform decision making in the distribution of resources
- *Subjects*: i.e., the researchers responsible for the document-related activities observed by the analysts

Actions

The main kinds of tasks involving choices among alternatives to be made by bibliometricians were summarized above, in the section on "A Conceptual Framework for Bibliometrics." These tasks include selection of the following:

- The unit type of evaluands
- A method of identifying the population of evaluands
- The variables used to characterize evaluands
- A method of operationalizing the variables

- A method of normalizing values
- A method of weighting the variables

The main kinds of decisions to be made by users of the results of bibliometric evaluations are:

- The level at which a given researcher, project, program, department, institution, etc., is to be funded or otherwise supported
- The formula or principle according to which available resources are to be distributed among the population of potential recipients

The main kinds of document-related choices to be made by researchers are:

- The frequency with which the researcher writes documents
- The coauthors with whom the researcher collaborates on a given document
- The order in which coauthors are listed on the document
- The topic(s) that the document is to cover
- The other documents that the document is to cite
- The venue(s) (e.g., the journal) to which the document is to be submitted

Values

The preeminent professional association for bibliometricians, the International Society for Scientometrics and Informetrics (ISSI), does not currently maintain a

code of ethics for reference by its members.[7] Candidates for the values and principles that are promoted by bibliometricians, policymakers, and researchers may instead be sought in the codes of ethics developed by standards-making bodies in closely related fields, such as evaluation, statistics, and publishing. For the present chapter, several such codes were mined with the aim of producing the lists of values and principles that follow:

- *Declaration on Professional Ethics* (International Statistical Institute [ISI], 2010)
- *Norms for Evaluation in the UN System* (United Nations Evaluation Group [UNEG], 2005a), *UNEG Ethical Guidelines for Evaluation* (UNEG, 2008), and *Standards for Evaluation in the UN System* (UNEG, 2005b)
- "Responsible Research Publication: International Standards for Authors" (Wager & Kleinert, 2011), and "Responsible Research Publication: International Standards for Editors" (Kleinert & Wager, 2011)
- *The European Code of Conduct for Research Integrity* (European Science Foundation [ESF], 2011)

Analysts

The kinds of values that typically appear in statements purporting to summarize the values held by professional statisticians and evaluators may be classified into

three broad groups, according to their status as characteristics of the products, methods, or agents of evaluative work.

Valued characteristics of the *products* (i.e., the outputs) of such work, such as rankings, include the following pair, each of which may be interpreted as a family of subproperties of varying significance:

- Quality (i.e., credibility, trustworthiness) of data: e.g.,
    - Accuracy
    - Completeness
    - Consistency
    - Absence of bias
- Fitness for purpose (i.e., utility, usefulness): e.g.,
    - Relevance
    - Timeliness
    - Accessibility
    - Clarity and transparency

The motivation for making the binary distinction drawn here is to highlight the difference between (a) final or intrinsic values, and (b) instrumental or extrinsic values. The usefulness of a given ranking can be determined only by external reference to the use to which it is put, whereas the credibility of a ranking can, at least in principle, be determined without reference to external purposes. In

the normal absence of a "ground truth" against which the product of an evaluative study may be compared, however, levels of data quality may be estimated in practice by examining the propensity of the methods selected by evaluators to produce outputs that are trustworthy. A breakdown of the valued characteristics of analysts' *methods* (i.e., processes) might proceed along the following lines:

- Fitness for purpose (i.e., propensity to produce outputs that are trustworthy and useful): e.g.,
    - Validity (i.e., extent to which methods are capable in practice of providing answers to the research questions to which they are applied)
    - Reliability (i.e., extent to which methods are capable in practice of providing reproducible results)

Valued characteristics (i.e., virtues) of analysts as *agents* include the following (commonly grouped under the family name of *integrity*):

- Impartiality
- Honesty
- Respectfulness (e.g., of rights)
- Accountability

Users

According to the codes examined, administrators and policymakers who make decisions informed by the results of bibliometric evaluations value characteristics of the *outcomes* of those decisions as follows:

- Cost-effectiveness (i.e., extent to which the benefits for the administrator of applying the results of the evaluation outweigh its costs for the administrator)
- Maximization of benefit-harm ratio (i.e., extent to which the combined benefits for members of all stakeholder groups, including evaluands, of applying the results of the evaluation outweigh the harms)

Valued characteristics of administrators' *methods* of applying the results of evaluative studies include the following:

- Fitness for purpose (i.e., propensity to produce outcomes that maximize welfare and cost-effectiveness): e.g.,
    - Fairness in distribution of reward (i.e., extent to which the resources distributed on the basis of the results of the evaluation are allocated in a manner demonstrated to be fair to recipients)
    - Transparency of purpose (i.e., extent to which administrators' goals, intentions, assumptions, and values are clarified)

The *virtues* of administrators and policymakers may be broken down in a similar way to that applied to analysts:

- Impartiality
- Honesty
- Respectfulness
- Accountability

Subjects

Values reported to be held by members of the research community may similarly be categorized in relation to outputs, methods, and agents; and, as before, intrinsic or final values may be distinguished from extrinsic or instrumental values that are defined relative to some external goal or purpose.

Valued characteristics of researchers' *outputs* are:

- Quality (i.e., credibility, trustworthiness) of work: e.g.,
    - Accuracy
    - Consistency
    - Completeness
    - Absence of bias
- Fitness for purpose (i.e., utility, usefulness): e.g.,
    - Relevance
    - Timeliness
    - Accessibility

- Clarity
- Completeness of documentation
- Impact

The last value mentioned—*impact*—may be treated roughly as equivalent to the "maximization of benefit-harm ratio" applied to administrators' outputs, above, since the particular kind of impact that is valued is positive impact. Impact on different groups may be valued to varying degrees, and a distinction is often drawn between impact on science or knowledge (i.e., impact within academia or the research sector) and impact on society.

Valued characteristics of researchers' *methods* include:

- Fitness for purpose (i.e., propensity to produce outputs that are trustworthy and useful): e.g.,
  - Validity (i.e., extent to which methods are capable in practice of providing answers to the research questions to which they are applied)
  - Reliability (i.e., extent to which methods are capable in practice of providing reproducible results)

Virtues of researchers as *agents* are:

- Impartiality: e.g.,

    - Impartiality in distribution of credit for prior work (i.e., extent to which all and only those works used by researchers are cited and/or acknowledged)
- Honesty: e.g.,
    - Honesty in submission (i.e., extent to which works submitted for publication are original, substantial, unique, genuine products of those claiming to be their authors)
- Respectfulness: e.g.,
    - Respectfulness of stakeholders' rights (i.e., extent to which the various rights of the members of all stakeholder groups are taken into account in the course of research)
- Accountability

Principles

Principles specifying the kinds of actions that have ethical warrant—that is, that are justifiable by reference to intentions or expected consequences that are intrinsically good—may be formulated by considering the kinds of decisions to be made in light of the values identified. The codes listed earlier provide some examples, of which a selection follows.

Analysts

Quality of Data

> [Statisticians should] strive to collect and analyze data of the highest quality possible. (ISI, 2010, p. 5)

Clarity and Transparency

> Evaluators should discuss, in a contextually appropriate way, those values, assumptions, theories, methods, results, and analyses that significantly affect the interpretation of the evaluative findings. (UNEG, 2005b, p. 17)

> [Statisticians should be] transparent about the statistical methodologies used and make these methodologies public. . . . In order to promote and preserve the confidence of the public, statisticians should ensure that they accurately and correctly describe their results, including the explanatory power of their data. It is incumbent upon statisticians to alert potential users of the results to the limits of their reliability and applicability. . . . Adequate information should be provided to the public to permit the methods, procedures, techniques, and findings to be assessed independently. (ISI, 2010, pp. 5–7)

Validity and Reliability

> Evaluation methodologies . . . should reflect the highest professional standards. . . . Evaluation processes [should] ensur[e] that evaluations are conducted in an objective, impartial, open and participatory [manner], based on empirically verified evidence that is valid and reliable, with results being made available. . . . The evaluation methodologies to be used for data collection, analysis and involvement of stakeholders should be appropriate to the subject to be evaluated, to ensure that the information collected is valid, reliable and sufficient to meet the evaluation objectives, and that the assessment is complete, fair and unbiased. . . . Evaluation methodologies should be sufficiently rigorous to assess the subject of evaluation and ensure a complete, fair and unbiased assessment. . . . Evaluation methods depend on the information sought, and the type of data being analysed. The data should come from a variety of sources to ensure its accuracy, validity and reliability, and that all affected people/stakeholders are considered. Methodology should explicitly address issues of gender and under-represented groups. (UNEG, 2005b, pp. 6, 13)

> [Evaluators should carry out] thorough inquiries, systematically employing appropriate methods and techniques to the highest technical standards, validating information using multiple

> measures and sources to guard against bias, and ensuring errors are corrected. (UNEG, 2008, p. 8)
>
> [Statisticians] are responsible for the fitness of data and of methods for the purpose at hand. . . . [They should] pursue promising new ideas and discard those demonstrated to be invalid . . . [and] work towards the logical coherence and empirical adequacy of . . . data and conclusions. (ISI, 2010, p. 5)

Impartiality

> Evaluators must ensure the honesty and integrity of the entire evaluation process. [Evaluators] also have an overriding responsibility to ensure that evaluation activities are independent, impartial and accurate. (UNEG, 2005b, p. 10)
>
> In carrying out his/her responsibilities, each statistician must be sensitive to the need to ensure that his/her actions are, first, consistent with the best interests of each group and, second, do not favor any group at the expense of any other. . . . [Statisticians should] use . . . statistical knowledge, data, and analyses for the Common Good to serve the society. . . . [Statisticians should] produce statistical results using . . . science and . . . not [be] influenced by pressure from politicians or funders. . . . [Statisticians should] strive to produce results that reflect the

> observed phenomena in an impartial manner. . . . Statisticians should pursue objectivity without fear or favor, only selecting and using methods designed to produce the most accurate results. . . . Available methods and procedures should be considered and an impartial assessment provided to the employer, client, or funder of the respective merits and limitations of alternatives, along with the proposed method. (ISI, 2010, pp. 4–6)

Respectfulness

> Evaluations [should be] carried out with due respect and regard to those being evaluated. . . . Evaluators should be sensitive to beliefs, manners and customs and act with integrity and honesty in their relationships with all stakeholders. . . . In line with the UN Universal Declaration of Human Rights and other human rights conventions, evaluators should operate in accordance with international values. . . . Evaluators should be aware of differences in culture, local customs, religious beliefs and practices, personal interaction and gender roles, disability, age and ethnicity, and be mindful of the potential implications of these differences when planning, carrying out and reporting on evaluations. . . . Evaluators should protect the anonymity and confidentiality of individual information. . . . The rights and well-being of individuals should

not be affected negatively in planning and carrying out an evaluation. (UNEG, 2005b, pp. 7, 10, 17)

Evaluations can have a negative effect on their objects or those who participate in them. Therefore evaluators shall seek to: minimize risks to, and burdens on, those participating in the evaluation; and seek to maximize the benefits and reduce any unnecessary harms that might occur from negative or critical evaluation, without compromising the integrity of the evaluation. (UNEG, 2008, p. 8)

[Statisticians should] respect the communities where data is collected and guard against harm coming to them by misuse of the results. . . . Findings should be communicated for the benefit of the widest possible community, yet attempt to ensure no harm to any population group. . . . In collaborating with colleagues and others in the same or other disciplines, it is necessary and important to ensure that the ethical principles of all participants are clear, understood, respected, and reflected in the undertaking. (ISI, 2010, pp. 5–7)

Users

Transparency of Purpose

> Make clear from the outset how the evaluation report will be used and disseminated. (UNEG, 2008, p. 11)

Respectfulness of Stakeholders' Rights

> Anticipate the different positions of various interest groups and minimize attempts to curtail the evaluation or bias or misapply the results. (UNEG, 2008, p. 11)

Subjects

Impartiality in Distribution of Credit for Prior Work

> Authors should represent the work of others accurately in citations and quotations. . . . Relevant previous work and publications, both by other researchers and the authors' own, should be properly acknowledged and referenced. The primary literature should be cited where possible. . . . Data, text, figures or ideas originated by other researchers should be properly acknowledged and should not be presented as if they were the authors' own. Original wording taken directly from publications by other researchers should appear

> in quotation marks with the appropriate citations. (Wager & Kleinert, 2011, p. 3)
>
> Important work and intellectual contributions of others that have influenced the reported research should be appropriately acknowledged. Related work should be correctly cited. References should be restricted to (paper or electronically) printed publications and publications "in print." (ESF, 2011, p. 14)

Honesty in Submission

> Work should not be submitted concurrently to more than one publication unless the editors have agreed to co-publication. . . . Authors should inform editors if findings have been published previously or if multiple reports or multiple analyses of a single data set are under consideration for publication elsewhere. Authors should provide copies of related publications or work submitted to other journals. . . . Multiple publications arising from a single research project should be clearly identified as such and the primary publication should be referenced. Translations and adaptations for different audiences should be clearly identified as such, should acknowledge the original source, and should respect relevant copyright conventions and permission requirements. (Wager & Kleinert, 2011, pp. 3–4)

[Authors should not engage in] repeated publication [or] salami-slicing. . . . Publication of the same (or substantial parts of the same) work in different journals is acceptable only with the consent of the editors of the journals and where proper reference is made to the first publication. In the author's CV such related articles must be mentioned as one item. (ESF, 2011, pp. 6, 14)

The authorship of research publications should accurately reflect individuals' contributions to the work and its reporting. . . . The criteria for authorship and acknowledgement should be agreed at the start of the project. Ideally, authorship criteria within a particular field should be agreed, published and consistently applied by research institutions, professional and academic societies, and funders. . . . Researchers should ensure that only those individuals who meet authorship criteria (i.e. made a substantial contribution to the work) are rewarded with authorship and that deserving authors are not omitted. Institutions and journal editors should encourage practices that prevent guest, gift, and ghost authorship. (Wager & Kleinert, 2011, pp. 1, 4)[8]

All authors, unless otherwise specified, should be fully responsible for the content of publication. Guest authorship and ghost authorship are not acceptable. The criteria for establishing the sequence of authors should be agreed by all, ideally at the start of

the project. Contributions by collaborators and assistants should be acknowledged, with their permission. (ESF, 2011, p. 7)

Editors should work to ensure that all published papers make a substantial new contribution to their field. Editors should discourage so-called "salami publications" (i.e., publication of the minimum publishable unit of research), avoid duplicate or redundant publication unless it is fully declared and acceptable to all (e.g., publication in a different language with cross-referencing), and encourage authors to place their work in the context of previous work (i.e., to state why this work was necessary/done, what this work adds or why a replication of previous work was required, and what readers should take away from it). (Kleinert & Wager, 2011, p. 5)

Editors should not attempt to inappropriately influence their journal's ranking by artificially increasing any journal metric. For example, it is inappropriate to demand that references to that journal's articles are included except for genuine scholarly reasons. In general, editors should ensure that papers are reviewed on purely scholarly grounds and that authors are not pressured to cite specific publications for non-scholarly reasons. (Kleinert & Wager, 2011, p. 3)

Omissions

Codes of ethics may themselves be evaluated along the lines developed above. Statements of normative principles are "fit for purpose" to the extent that they are accessible to, and considered relevant by, the members of their intended audience. Different methods, of course, would be needed if we wished to conduct a sociological study of the values that individuals actually claim to hold, and thus of the degree to which researchers, bibliometricians, and administrators are guided, in practice, by the norms long codified by their professional associations. Meanwhile, we can proceed by pointing to gaps in the codes, where guidance on certain specifics would be especially useful, yet is unfortunately absent.

The most significant omission is that of a principle of *distributive justice*. Existing statements of principles are largely silent on the issue of the right way to allocate rewards to researchers. The general question addressed by theories of distributive justice is this: On the basis of what principle should benefits and burdens of any kinds (including economic and cultural goods and services) be distributed among populations of recipients? Justice, or fairness, is the label conventionally given to the valued property of distributions of benefits. Different theories of distributive justice provide justifications for different principles by which that value may be maximized (see, e.g., Cozzens, 2007; Lamont & Favor, 2007). For example: Principles of *strict equality* define fairness as the extent to

which every member of a population receives the same quantity of net benefits. No characteristics of individual recipients are relevant to distributions based on strict equality; principles of *relative equality* specify some particular characteristic of recipients (such as need, desert, or status) in accordance with which benefits should be distributed. Yet other principles allow for inequalities only to the extent that the least advantaged are better off than they would be under strict equality. *Libertarian* theories deny the primacy of equality as a value, and consider distributions to be fair to the extent that certain freedoms and rights of recipients are respected.

A prevailing, if frequently left unstated, assumption held by participants in the national and international governance of research is that resources should be distributed in accordance with *desert* (i.e., the extent to which recipients are deserving of reward). The absence of an attendant justification for this general principle is less problematic than is the (equally understandable) absence of guidance in dealing with measurement issues of four related kinds that are perennials for distributive-justice theorists and evaluation theorists alike:

- On what basis should we determine which *elementary characteristics* (e.g., merit, need; past performance, future potential; intrinsic quality, extrinsic impact) are to be included in the calculus of overall desert?

- On what basis should we determine how these various variables are to be *operationalized* in forms (e.g., publication counts, citation counts) that are measurable?
- On what basis should we determine how the values of these various variables are to be *normalized* (e.g., by time frame)?
- On what basis should we determine how these various variables are to be *weighted*?

In general, the need is for a rational principle for determining the right way of measuring amounts of desert. Existing codes of ethics lack advocacy of any such principle, and policymakers' and analysts' selections of relevant characteristics, and of methods of operationalization, normalization, and weighting, tend to be made on a largely ad hoc basis.

Violations

The technical and methodological problems faced by evaluative bibliometricians are numerous and widely discussed, and their effects on the validity and reliability of bibliometric methods are relatively well understood (see, e.g., Bornmann, Metz, Neuhaus, & Daniel, 2008; Moed, 2007; Pendlebury, 2009; Sivertsen, 1997). For example:

- The databases of publications and citations on which counts are based tend to contain errors that are not always distributed uniformly, and tend to lack unbiased coverage of all types of citing documents (e.g., books as well as journal articles), all languages, and all fields. Conclusion: The use of counts derived from such databases to compare authors whose oeuvres are well covered with those whose oeuvres are not is invalid.
- Citation counts are not distributed uniformly or normally across cited documents; rather, the distributions are heavily skewed, with the result that mean counts work poorly as descriptions. Conclusion: The use of metrics based on mean counts (e.g., impact factors) as proxies for individual counts is invalid.
- Documents in some disciplines tend to attract higher citation counts simply because those disciplines are large or highly productive. Conclusion: The use of citation counts to compare evaluands across disciplines is invalid.
- Many authors base their decisions to cite a given document on reasons other than whether they have used it or not. Conclusion: The use of citation counts as evidence of impact is invalid.

Should we remain in any doubt about the desirability of using methods whose validity has already been shown to be suspect, we are now in a position to ask serious questions about the intrinsic value of such methods, on the basis of

our observation of the absence of justification for (or of systematic compliance with) principles of distributive justice. Is it *fair* to compare citation counts and impact factors without normalizing for disciplinary differences? If such normalization is required, what are the relevant dimensions of difference, and at what level of aggregation (discipline, field, area, individual researcher) should the normalization take place? Is it *fair* to treat productivity and impact as indicators of research quality? Is it *fair* to treat publication counts as evidence of productivity, and citation counts as evidence of impact?

Meanwhile, much is made in the codes of the supposed moral unacceptability of the various kinds of activities in which candidates for reward (rather than its distributors) engage with the aim of "gaming" the system, by corrupting indicators so that they can no longer be treated as valid measures of desert:

- The "salami-slicing" strategy, by which authors divide up their research results for separate publication in a series of "least publishable units"
- The "repeated-publication" strategy, by which authors submit very similar papers to multiple venues
- The "guest-author" strategy, by which those who did not contribute to a publication nevertheless claim authorship of it

- The "citation trawling" strategy, by which journal editors encourage (or even require) authors of submissions to cite their journals

(Nothing is said in the codes about the acceptability of the similarly motivated institutional practice of hiring, and paying large salaries to, highly cited scholars in order to boost institutional counts in advance of national research assessment exercises.)

It might be argued that such instances of "gaming" are quite rational reactions to the perception that one is being forced to participate in a system of evaluation that is unfair to begin with (see, e.g., Frey & Osterloh, 2011). Administrators typically have reasons of two good kinds for using bibliometric techniques in evaluations:

- It is *possible* to distribute reward on the basis (at least partially) of quantitative measurement of the frequency of occurrence and/or strength of document-related events (publications, citations, etc.).
- Distributing reward on the basis of quantitative measurement is more *cost-effective* than doing so on the basis of qualitative peer review, which requires hard work over long periods by experts on the topics of a wide range of publications.

A far greater challenge for administrators is to demonstrate the intrinsic *fairness* of the quantitative approach. With such a challenge in mind, the framework

presented in this chapter is intended for use in identifying the issues requiring attention, in reaching an understanding of the reasons for bibliometricians' past disinclination to adopt a code of professional ethics, and ultimately in exerting appropriate levels of pressure, on groups and institutions with authority and influence in the field, to require their members routinely to provide justifications *on ethical grounds* of their decisions, actions, and practices.

Conclusion

In this chapter, I have focused on the ethical implications of using evaluative bibliometrics to inform decision making in the distribution of reward. The self-imposed limitation, whereby consideration of applications of bibliometric techniques to information retrieval (IR) was excluded, is quite arbitrary. Other uses of bibliometrics are no less fraught with ethical issues. One challenge for the designers of search engines that make recommendations of documents in accordance with counts of prior usage events (links, views, downloads, etc.) is to develop a convincing response to the charge that the *Matthew effect* (see, e.g., Rigney, 2010)—an ever-increasing inequality between higher-ranked and lower-ranked documents resulting from the tendency of higher-ranked documents to attract more usage—is the product of a mechanism that distributes rank unfairly. Any context in which a "rich-get-richer" phenomenon of cumulative advantage is

observed would certainly seem to be a prime candidate for justice-theoretic analysis. It is hoped that the framework presented in this chapter may help to stimulate further work in this area.

Notes

{Notes_begin}

1. *Value* is a term that, potentially confusingly, has at least three distinct senses: (1) an amount, quantity, or number serving as a measurement of the extent or degree to which (or the level or rate at which) any phenomenon exhibits a given property; (2) any kind of state, condition, property, etc., attributable to agents, objects, events, or other phenomena, as individuals or in aggregations, that is held by some agent (or group of agents) to be good; and (3) (as here) the amount or quantity of goodness intrinsic to, or potentially generated by, an evaluand. On different occasions in this chapter, different senses are intended; it is hoped that the context makes the intention clear in each case.

2. See previous note. Here *value* is used in sense 1. (The sense of *variable* here is close to that of sense 2.).

3. Some indicators are the products of analysis of a hybrid form. The *h*-index, for example, is a measure in which publication counts and citation counts are combined.

4. That is, the extent to which researchers are deserving, or worthy of receiving reward.

5. Core journals in which such contributions are published include the *Journal of Informetrics*, the *Journal of the American Society for Information Science and Technology*, *Research Evaluation*, and *Scientometrics*.

6. A useful resource in this context is the Center for the Study of Ethics in the Professions (CSEP) at the Illinois Institute of Technology (IIT). According to its website (http://ethics.iit.edu/about/history-mission-center, ¶ 1), CSEP was established in 1976 "to promote research and teaching on practical moral problems in the professions." It is "the first interdisciplinary center for ethics to focus on the professions," and "one of the nation's leading centers for practical and professional ethics." CSEP maintains an online collection of over 850 codes of ethics.

7. According to its website (http://www.issi-society.info/mission.html), ISSI was established in 1993 with the aims "to encourage communication and exchange of

professional information in the field of scientometrics and informetrics, to improve standards, theory, and practice in all areas of the discipline; to stimulate research, education, and training, and to enhance the public perception of the discipline."

8. "Guest authors are those who do not meet accepted authorship criteria but are listed because of their seniority, reputation or supposed influence; gift authors are those who do not meet accepted authorship criteria but are listed as a personal favour or in return for payment; ghost authors are those who meet authorship criteria but are not listed" (Wager & Kleinert, 2011, p. 4). Reliable measurements of the prevalence of guests, gifts, and ghosts are hard to come by. One might reasonably expect to see substantial disciplinary differences, both in the frequencies of occurrence of these quasi-authorial acts, and in administrators' and scholars' perceptions of the demerits of such acts. For example, there is anecdotal evidence to suggest that, in medicine and some related fields, ghost authorship is a practice that is both relatively common and generally perceived as benign.

{Notes_end}

References

<jrn>Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century: A review. *Journal of Informetrics*, *2*(1), 1–52.</jrn>

<edb>Bar-Ilan, J. (2010). Informetrics. In M. J. Bates & M. N. Maack (Eds.), *Encyclopedia of library and information sciences* (3rd ed., pp. 2755–2764). Boca Raton, FL: CRC Press.</edb>

<jrn>Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science & Technology*, *36*, 3–72.</jrn>

<jrn>Bornmann, L., Metz, R., Neuhaus, C., & Daniel, H.-D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics*, *8*, 93–102.</jrn>

<jrn>Cozzens, S. E. (2007). Distributive justice in science and technology policy. *Science & Public Policy*, *34*(2), 85–94.</jrn>

<eref>European Science Foundation. (2011). *The European code of conduct for research integrity*. Strasbourg, France: European Science Foundation. Retrieved from

http://www.esf.org/fileadmin/Public_documents/Publications/Code_Conduct_ResearchIntegrity.pdf. </eref>

<other>Frey, B. S., & Osterloh, M. (2011). *Ranking games* (CREMA Working Paper No. 11). Basel, Switzerland: Center for Research in Economics, Management and the Arts.</other>

<eref>International Statistical Institute. (2010). *Declaration on professional ethics*. The Hague, The Netherlands: International Statistical Institute. Retrieved from http://www.isi-web.org/images/about/Declaration-EN2010.pdf</eref>

<eref>Kleinert, S., & Wager, E. (2011). Responsible research publication: International standards for editors: A position statement developed at the 2nd World Conference on Research Integrity, Singapore, July 22–24, 2010. In T. Mayer & N. Steneck (Eds.), *Promoting research integrity in a global environment* (pp. 317–328). Singapore: Imperial College Press / World Scientific Publishing. Retrieved from http://publicationethics.org/international-standards-editors-and-authors.</eref>

<eref>Lamont, J., & Favor, C. (2007). Distributive justice. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. Stanford, CA: Stanford University. Retrieved from http://plato.stanford.edu/entries/justice-distributive/</eref>

<jrn>Moed, H. F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science & Public Policy*, *34*(8), 575–583.</jrn>

<bok>Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill, NJ: Computer Horizons.</bok>

<jrn>Pendlebury, D. A. (2009). The use and misuse of journal metrics and other citation indicators. *Archivum Immunologiae et Therapiae Experimentalis*, *57*, 1–11.</jrn>

<bok>Rigney, D. (2010). *The Matthew effect: How advantage begets further advantage*. New York: Columbia University Press.</bok>

<bok>Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage.</bok>

<bok>Shafer-Landau, R. (2010). *The fundamentals of ethics*. New York: Oxford University Press.</bok>

<edb>Sivertsen, G. (1997). Ethical and political aspects of using and interpreting quantitative indicators. In M. S. Frankel & J. Cave (Eds.), *Evaluating science and scientists: An East-West dialogue on research evaluation in post-communist Europe* (pp. 212–220). Budapest, Hungary: Central European University Press.</edb>

<jrn>Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information Processing & Management*, *28*, 1–3.</jrn>

<eref>United Nations Evaluation Group. (2005a). *Norms for evaluation in the UN system*. New York: United Nations Evaluation Group. Retrieved from http://www.uneval.org/normsandstandards.</eref>

<eref>United Nations Evaluation Group. (2005b). *Standards for evaluation in the UN system*. New York: United Nations Evaluation Group. Retrieved from http://www.uneval.org/normsandstandards.</eref>

<eref>United Nations Evaluation Group. (2008). *UNEG ethical guidelines for evaluation*. New York: United Nations Evaluation Group. Retrieved from http://www.unevaluation.org/ethicalguidelines.</eref>

<eref>Wager, E., & Kleinert, S. (2011). Responsible research publication: International standards for authors: A position statement developed at the 2nd World Conference on Research Integrity, Singapore, July 22–24,

2010. In T. Mayer & N. Steneck (Eds.), *Promoting research integrity in a global environment* (pp. 309–316). Singapore: Imperial College Press / World Scientific Publishing. Retrieved from http://publicationethics.org/international-standards-editors-and-authors.</eref>

<edb> Whitley, R., & Gläser, J. (Eds.). (2007). *The changing governance of the sciences: The advent of research evaluation systems*. Dordrecht, The Netherlands: Springer.</edb>